On the Interaction of Information and Decisions in Dynamic Networked Systems

by

Yi Ouyang

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Electrical Engineering: Systems)
in The University of Michigan
2016

Doctoral Committee:

Professor Demosthenis Teneketzis, Chair
Professor Tamer Basar, University of Illinois at Urbana-Champaign
Professor Mingyan Liu
Assistant Professor Ashutosh Nayyar, University of Southern California
Professor Mark P. Van Oyen

ACKNOWLEDGEMENTS

I would like to sincerely express my gratitude to my advisor Professor Demosthenis Teneketzis for guiding me to a research area I enjoy. When I entered graduate school, I wanted to study fundamental problems in multi-agent communication networks, but I knew nothing about stochastic control which ends up being my research focus. I am extremely fortunate to have an advisor who provides continuous support and excellent guidance through the years of my Ph.D. study and research. I would like to thank him for all his patient and stimulating discussions, and all the time he spent to help me improve my writings and presentations.

I would also like to thank all my committee members, Professor Tamer Basar, Professor Mingyan Liu, Professor Ashutosh Nayyar, and Professor Mark Van Oyen. Their insightful comments and constructive suggestions are essential for me to complete my Ph.D. thesis.

I am thankful to the opportunity to work with many friends at Michigan. In particular, I would like to thank Yang Liu, Erik Miehling, Mohammad Rasouli, and Hamidreza Tavafoghi for many enjoyable discussions.

I am also grateful to all my friends in Ann Arbor for all the time we spent together having potlucks, playing boardgames, playing in bridge tournaments, exploring the nature, and chatting on everything. These precious moments constitute a major part of my live in Ann Arbor.

Finally, I would like to thank Chi-Mei and my family for their love, and their unconditional support to all my decisions.

TABLE OF CONTENTS

LIST OF FIGURES

Figure

LIST OF APPENDICES

Appendix

ABSTRACT

On the Interaction of Information and Decisions in Dynamic Networked Systems

by

Yi Ouyang

Chair: Demosthenis Teneketzis

Efficient operation of modern dynamic networked systems, such as communication systems, queueing networks, power systems, and surveillance systems, can significantly improve our quality of life. The operation of a dynamic networked system involves series of decision making processes by many decision makers (DMs) who may or may not have the same information, and may or may not share the same objective.

The quality of each DM's decision depends on the quality of the information available for decision-making in the network. Since the network is dynamic, the information available to the DMs over time is a dynamic process that depends on the DMs' decision rules. Information affects decisions, and decisions influence information. This interaction between information and decisions in dynamic networks results in complex decision-making problems.

In this thesis, we study the impact of the information-decision interaction on system performance within the context of: (i) centralized stochastic control; (ii) decentralized stochastic control; and (iii) game theory. Specifically, within the context of centralized stochastic control, we study a multi-state channel sensing problem,

and discover sets of conditions sufficient to guarantee the optimality of a myopic policy. Within the context of decentralized stochastic control, we consider a decentralized routing problem as well as a multiple access communication problem; we discover an optimal decentralized routing policy for the routing problem, and an efficient decentralized multiple access protocol. Within the context of game theory, we study a general model of dynamic stochastic games with asymmetric information; we introduce the concept of common information based perfect Bayesian equilibrium (CIB-PBE), and provide a sequential decomposition for the dynamic games that leads to an algorithm to determine CIB-PBE.

CHAPTER I

Introduction

1.1 Background

Many modern social-technological systems such as communication systems, queueing networks, power systems, surveillance systems and social networks are dynamic networked systems. Efficient operation of such networked systems can significantly improve our quality of life. The operation of a dynamic network involves a series of decision making processes by one decision maker (DM) or multiple DMs. For example, data communication in a communication system depends on the users' transmission strategies; a queueing network's operation is based on the scheduling and routing strategies of the routers; a surveillance system's performance depends on the communication and control strategies of sensors and controllers.

A key challenge in the design of efficient decision strategies is the presence of uncertainty in the network. The uncertainty includes the randomness in the network dynamics, in the measurements/observations, and in the human behavior. In the presence of uncertainty, the quality of each DM's decision depends on the quality of information available for decision-making; more accurate information about the uncertain network leads to higher quality of decisions. In a dynamic network, each DM collects information generated throughout the network from his own measurements as well as the actions/signals generated by other DMs. Since the network is

dynamic, information generation is a dynamic process that depends on the DMs' control/decision strategies over time. For example, the information available to a surveillance camera depends on its direction, and the direction is determined by the camera's rotation strategy. Therefore, the quality of the DMs' future information is affected by the DMs' current decisions which themselves are influenced by their current information. This interaction between information and decisions determines the operation of a dynamic networked system. Research on the interaction between information and decisions plays a key role in determining efficient designs of modern social-technological systems.

In this thesis, we study the impact of the information-decision interaction on a networked system's performance. Since the interaction between information and decisions has different features under different information structures and different models on DMs' behavior, we introduce below the concept of information structure along with the model on DMs' behavior.

1.1.1 Information Structure

The information structure of a dynamic networked system specifies the information that is available at each time to each DM for decision-making purposes. Information structures can be classified into centralized and decentralized.

1.1.1.1 Centralized Information Structure

A dynamic network has centralized information structure if there is either one DM or multiple DMs that share the same information at every time. Furthermore, the DMs have perfect recall so that at each time they remember all the information available at previous times. Centralized information structure arises in many classical networked systems where a centralized coordinator collects all information in the network. It also arises in small systems with zero-delay communication among DMs.

In a networked system with centralized information structure, the DMs' decisions affect the system's evolution and its performance, and determine the quality of the information that will be collected in the future. Consequently, an efficient decision strategy needs to consider two factors: (a) its effect on the network dynamics and the system's performance, and (b) its effect on future information collection. This dual function of a decision strategy is called dual-control in the stochastic control literature [1].

1.1.1.2 Decentralized Information Structure

A dynamic networked system has decentralized information structure if its information is not centralized (see [2] for a comprehensive discussion of decentralized information structure). As the size of modern networks grows, it may be impossible for a centralized coordinator to collect all information in a large scale network; it is also unrealistic to assume zero-delay communication among DMs. Therefore, understanding the operation of modern dynamic networked systems requires the analysis of systems with decentralized information structure.

In contrast to centralized information, each DM in a networked system with decentralized information structure has different information about the system from other DMs. Since a DM's decision is selected based on his available information, this decision implicitly reveals part of the DM's information. By observing a DM's decision (or the effect of the decision on the system), other DMs can infer part of that DM's information. Such a phenomenon is called signaling in decentralized control [3] and in signaling games [4].

Similar to centralized networks, a DM's decision in a decentralized network affects the system's evolution, its performance, and the DM's future information collection. Furthermore, in the presence of signaling, each DM's decision also affects other DM's future information collection. As a result, in a networked system with decentral-

ized information structure, the selection of a DM's decision strategy should consider three factors: (a) the strategy's effect on the network dynamics and the system's performance; (b) its effect on the DM's future information collection; (c) its effect on the other DM's future information collection. Thus, in systems with decentralized information, a decision strategy has a triple function.

1.1.2 Model on DMs' Behavior

In a dynamic networked system, there are two different models on DMs' behavior depending on their objectives. Either the DMs cooperatively accomplish a common objective (non-strategic), or they attempt to achieve their own objectives through strategic interaction (strategic).

1.1.2.1 Non-Strategic Behavior

When the DMs in a dynamic networked system are non-strategic, they share the same objective; their decision-making problem is to design decision strategies that optimize the system's performance. Such problems are also called team problems (see [3]). Team problems arise in many networked control systems where each DM is a non-strategic controller.

In the team situation, the DMs cooperate with each other to improve the quality of their information as well as to enhance the performance of the system. As a result, the DMs' strategies (functional forms, not the realized decisions) are commonly known by all of them because they can all agree on the strategies before the system begins to operate. This means that non-strategic DMs have a consistent view of the interaction between information and decisions in the system.

1.1.2.2 Strategic Behavior

When the DMs in a dynamic networked system are strategic, they attempt to achieve their own objectives through strategic interaction. The DMs' strategic behavior results in a dynamic stochastic game. Game problems arise in systems where the DMs are strategic agents having different (partially conflicting) interests.

In this situation, a DM will attempt to hide information from his competitors, and reveal part of it only when this revelation improves his utility. Consequently, the strategies of strategic DMs are not commonly known among them. Therefore, each DM needs to predict other DMs' strategies to make decisions. Since each DM knows that all the DMs maximize their objectives, any reasonable prediction about a DM's strategy should be constructed based on the following conditions: (i) the DM maximizes his own objective under his predictions, and (ii) the DM's predictions about other DMs' strategies are also constructed based on (i) and (ii). Such circular construction of predictions leads to the concept of *Bayesian Nash equilibrium* (BNE) (see [5, 6]). A BNE is a collection of predictions about the DMs' strategies, such that, each DM's strategy in the BNE maximizes his objective when he uses the BNE to predict other DMs' strategies. BNE is the proper solution concept to analyze stochastic games because a BNE consists of reasonable predictions that satisfy (i) and (ii). The networked system will operate according to a BNE when each DM uses his BNE strategy along with the BNE predictions about other DMs' strategies. Then the BNE strategies are commonly known by all DMs, and information will interact with decisions based on the BNE strategies.

1.2 Decision-Making Problems Investigated in the Thesis

Depending on the information structure and the model on DMs' behavior of a dynamic networked system, there are four classes of decision-making problems described

by the following table.

	Centralized	Decentralized
Non-Strategic	Centralized Stochastic Control	Decentralized Stochastic Control
Strategic	Dynamic Stochastic Games with Symmetric Information	Dynamic Stochastic Games with Asymmetric Information

Below we discuss issues on the interaction of information and decisions associated with the four classes of problems defined in the table above. In this thesis we focus on the three classes of problems: (i) centralized stochastic control; (ii) decentralized stochastic control; and (iii) dynamic stochastic games with asymmetric information. For dynamic stochastic games with symmetric information, we refer the interested reader to [5–9] and references therein.

1.2.1 Centralized Stochastic Control

In a centralized networked system with non-strategic DMs, all the cooperative DMs can be viewed as a centralized DM who selects decisions to optimize the system's performance. The centralized DM's decision-making problem results in a centralized stochastic control problem. Centralized stochastic control can be formulated in the general framework of partially observed Markov decision processes (POMDPs) [1]. In POMDP, the DM's decision depends on the *information state* that summarizes the DM's available information. The information state belongs to a high-dimensional space in general. Because of the high-dimensional space, solving a general POMDP numerically has very high complexity (PSPACE-complete [10]). Such high complexity is referred to as the *curse of dimensionality*.

One approach to dealing with the curse of dimensionality in POMDP is to discover properties for the information states. For that matter, we investigate a general multi-state channel sensing problem.

In the channel sensing problem, a DM has access to a communication network consisting of multiple channels. Each channel is modeled as a Markov chain. At each time instant the DM selects one channel, observes its state, and uses it to transmit data. A reward depending on the state of the selected channel is obtained for each transmission. The objective is to design a channel sensing policy that maximizes the expected total reward collected over a finite or infinite horizon. The above channel sensing problem arises in opportunistic scheduling over fading channels and cognitive radio networks [11]. In this channel sensing problem, in addition to transmission, the decision/selection allows the DM to observe the quality of the selected channel. Therefore, different selections provide different information about the states of the channels; hence, different decisions result in different information states for the DM. We study the effect of decisions on information states for the DM, and discover properties that allow us to order information states under certain conditions. Such properties lead to the optimality of a myopic sensing policy.

1.2.2 Decentralized Stochastic Control

In a decentralized networked system with non-strategic DMs, multiple DMs cooperatively optimize the system's performance. The design of DMs' optimal strategies results in a decentralized stochastic control problem. In a decentralized stochastic control problem, a DM's information is not directly observed by other DMs. Nevertheless, when a DM's strategy depends on his information, part of his information may be revealed/transmitted through his actions. This phenomenon is referred to as signaling as discussed in Section 1.1.1.2. Using a signaling strategy, a DM can transmit part of his information to improve the quality of some other DM's future information. When signaling occurs, each DM's decision will affect both his future information as well as the future information of all other DMs in the network. Therefore, signaling complicates the interaction between information and decisions, and makes

the design of efficient strategies conceptually and computationally challenging. The computational complexity to solve a general decentralized stochastic control problem is NEXP-hard [12].

Recently, the *common information approach* to decentralized stochastic control, proposed in [13, 14], addresses some of the conceptual difficulties arising from signaling. When all DMs have perfect recall, the DMs' *common information* at a time includes the information available to all DMs at that time. Each DM's *private information* at a time includes his information that is not common information. The common information approach uses the DMs' common information to coordinate their strategies by identifying an information state sufficient for performance evaluation (see [15]). This information state is a common belief on the system state and the DMs' private information based on the common information. The common belief allows the DMs to consistently assess the status of the system and to consistently predict how the DMs use their private information. Thus, the common information approach allows us to focus on the interaction of the common belief and decisions when signaling occurs. We show how the this approach works in a decentralized routing problem and a multiple access communication problem.

1.2.2.1 Decentralized Routing

Routing problems arise in many modern technological systems such as communication networks, transportation networks and sensor networks. There are many results on centralized routing in networked systems. However, very few results on optimal routing under decentralized information are currently available.

In this thesis, we consider a decentralized routing problem in a system consisting of two service stations in parallel and two DMs, each DM is affiliated with one service station. Each station has an infinite size queue with Bernoulli arrivals and departures. The network is informationally decentralized: each DM only knows perfectly the

information about its own station. At any time, a DM can route one of the customers waiting in its own station to the other station. At each time a holding cost is incurred at each station due to the customers waiting at that station. The objective is to determine routing strategies for the DMs that minimize the average cost per unit time over a finite horizon or infinite horizon.

In this routing network, signaling occurs between the two DMs through their routing decisions; whenever a DM decides to send or not to send a customer from its own station to the other station it communicates partial information about its queue length to the other station. For example, by receiving a customer from the other station, a DM may infer that the queue length at the other station is above a pre-specified threshold. This information from signaling may allow the DM to have a better estimate of the queueing system. Based on how signaling affects the DMs' common beliefs (the information state constructed from their common information), we explicitly determine optimal decentralized routing policies for the DMs.

1.2.2.2 Multiple Access Communication

Multiple access communication has played a crucial role in the operation of many communication systems, including satellite networks, radio networks, wired and wireless Local Area Networks (LANs). One important feature of multiple access is its decentralized information structure. Consequently, when multiple users share a common communication channel, coordination among them is essential to resolve collision issues. The design of efficient coordination mechanisms/protocols is a challenging problem.

We consider a slotted multiple access system where multiple DMs share a common collision channel. Each DM is equipped with an infinite size buffer and observes Bernoulli arrivals to its own queue. In addition to their local information, all DMs receive a common feedback from the channel. The feedback indicates whether the

previous transmission was idle, successful or collision.

Using signaling through the DMs' common beliefs (the information state constructed from the common feedback), we design a *common information-based multiple access protocol* (CIMA). In CIMA, each DM constructs upper bounds on the lengths of the queues of all DMs, including himself, based on previous transmission strategies and the common feedback. Since the upper bounds are common knowledge, the DMs can coordinate their transmission to avoid collision through the common upper bounds. We prove that without knowledge of any statistics, CIMA achieves the full throughput region of the collision channel and an average delay that is linear in the number of the DMs.

1.2.3 Dynamic Stochastic Games with Symmetric Information

Dynamic centralized networked systems with strategic DMs result in dynamic stochastic games with symmetric information. In these games, all the DMs share the same information but have different objectives. Therefore, each DM makes decisions anticipating other DMs' strategies. An appropriate solution concept for this class of games is Markov perfect equilibrium (MPE) [16], a refinement of BNE. At each stage of a dynamic game with symmetric information, there is an information state (Markov state) that can summarize the DMs' history of information. The information state can be utilized to provide a sequential decomposition of the dynamic game, and MPE can be computed through backward induction.

1.2.4 Dynamic Stochastic Games with Asymmetric Information

Dynamic decentralized networked systems with strategic DMs result in dynamic stochastic games with asymmetric information. In such problems, the DMs have different objectives and different information. Therefore, each DM needs to anticipate the other agents' strategies and to form beliefs about the other DMs' private infor-

mation. *Perfect Bayesian equilibrium* (PBE), a refinement of BNE, is an appropriate solution concept for this class of games. A PBE consists of a pair of strategy profile and a belief system for all DMs that jointly must satisfy *sequential rationality* and *consistency* [5, 6]. That is, every DM's strategy should be a best response under the belief system, and the belief system should be consistent with the DMs' strategies when signaling occurs (see the discussion of signaling in Section 1.1.1.2). This circular dependence between information (belief system) and decisions (strategy) makes it difficult to compute PBE. As a result, sequential computation of equilibria for stochastic dynamic games with asymmetric information is available only for special instances (see [14, 17–26] and references therein).

In this thesis, we study the interaction between information and decisions in a general model of dynamic stochastic games with asymmetric information. We identify an information state from the DMs' common information using ideas from the common information approach described in Section 1.2.2 for decentralized stochastic control. Using the information state, we introduce a subclass of PBE called *common information based perfect Bayesian equilibria* (CIB-PBE) and provide a sequential decomposition for the dynamic game. Such a decomposition leads to a backward induction algorithm to compute CIB-PBE. CIB-PBE and the associated decomposition resembles MPE for dynamic games with symmetric information.

1.3 Organization and Contributions of the Thesis

1.3.1 Chapter II-Centralized Stochastic Control: Multi-State Channel Sensing

We investigate the channel sensing problem in Chapter II. We generalize the "positively correlated" condition for two-state channels (see [27]) to multi-state channels and discover sets of conditions sufficient to guarantee the optimality of a myopic sens-

ing policy. We also develop a MDP approximation for the multi-state channel sensing problem resulting in a dynamic programming algorithm that is computationally feasible for a small number of channels. In addition, we consider the situation where the DM can select multiple channels at each time. We compare the performance of the myopic policy and the near-optimal policy generated from the MDP approximation when multiple selections are available.

1.3.2 Chapter III-Decentralized Stochastic Control-Part I: Decentralized Routing

We formulate a routing problem with decentralized information structure in Chapter III. We show that an optimal decentralized strategy is described by a single threshold routing policy where the threshold depends on the common information between the two DMs/controllers. We explicitly determine this threshold. For the case of decentralized routing with bursty arrivals, we construct a decentralized routing policy described by multiple thresholds that are functions of the common information. When bursty arrivals are finite, the decentralized routing policy can balance the queueing system.

1.3.3 Chapter IV-Decentralized Stochastic Control-Part II: Multiple Access Communication

We study the multiple access communication problem in Chapter IV. We present a common information-based multiple access protocol (CIMA) that has the following features: it is collision-free, it achieves full throughput and average delay that is linear in the number of DMs. The CIMA protocol is simple to implement, as at each time instant it only requires knowledge of the upper bounds on each DM's queue length. The upper bounds on the DMs' queue lengths are common knowledge and are updated in a simple manner.

1.3.4 Chapter V-Dynamic Stochastic Games with Asymmetric Information

We consider a general model of dynamic stochastic games with asymmetric information in Chapter V. The key contributions of Chapter V are: (1) The introduction of a subclass of PBE called common information based perfect Bayesian equilibria (CIB-PBE) for dynamic games with asymmetric information. A CIB-PBE consists of a pair of strategy profile and a belief system that are sequentially rational and consistent. (2) The sequential decomposition of stochastic dynamic games with the asymmetric information through an appropriate choice of information state. This decomposition provides a backward induction algorithm to find CIB-PBE for dynamic games where signaling occurs. The decomposition and the algorithm are illustrated by an example from multiple access communication. (3) The proof of existence of CIB-PBE for a subclass of stochastic dynamic games with asymmetric information.

1.3.5 Chapter VI-Conclusion and Future Directions

We conclude in Chapter VI and provide some thoughts on future directions that could extend the solution methods and results presented in this thesis.

1.4 Notation

Random variables are denoted by upper case letters, their realization by the corresponding lower case letter. In general, subscripts are used as time index while superscripts are used to index users/agents/controllers/stations. For time indices $t_1 \leq t_2$, $X_{t_1:t_2}$ (resp. $f_{t_1:t_2}(\cdot)$) is the short hand notation for the variables $(X_{t_1}, X_{t_1+1}, ..., X_{t_2})$ (resp. functions $(f_{t_1}(\cdot), \ldots, f_{t_2}(\cdot))$). When we consider the variables (resp. functions) for all time, we drop the subscript and use X to denote $X_{1:T}$ (resp. $f(\cdot)$ to denote $f_{1:T}(\cdot)$). For variables $X_t^1, \ldots, X_t^N$ (resp. functions $f_t^1(\cdot), \ldots, f_t^N(\cdot)$), we use

$X_t := (X_t^1, \ldots, X_t^N)$ (resp. $f_t(\cdot) := (f_t^1(\cdot), \ldots, f_t^N(\cdot))$) to denote the vector of the set of variables (resp. functions) at t, and $X_t^{-n} := (X_t^1, \ldots, X_t^{n-1}, X_t^{n+1}, \ldots, X_t^N)$ (resp. $f_t^{-n}(\cdot) := (f_t^1(\cdot), \ldots, f_t^{n-1}(\cdot), f_t^{n+1}(\cdot), \ldots, f_t^N(\cdot))$) to denote all the variables (resp. functions) at t except the one indexed by n. For a policy/strategy/protocol g, we use X^g to indicate that the random variable X^g depends on the choice of g. $\mathbb{P}(\cdot)$ and $\mathbb{E}(\cdot)$ denote the probability and expectation of an event and a random variable, respectively. For a set $\mathcal{X}$, $\Delta(\mathcal{X})$ denotes the set of all beliefs/distributions/PMFs (Probability Mass Functions) on $\mathcal{X}$. We use vectors in $\mathbb{R}^{\mathbb{Z}_+}$ to denote PMFs in $\Delta(\mathbb{Z}_+)$, where $\mathbb{Z}_+$ denotes the set of non-negative integers. We also use a constant in $\mathbb{Z}_+$ to denote the corner PMF that represents a constant r.v.. i.e. a constant $c \in \mathbb{Z}_+$ denotes the PMF whose entries are all zero except the cth. For random variables X, Y with realizations x, y, $\mathbb{P}(x|y) := \mathbb{P}(X = x|Y = y)$ and $\mathbb{E}(X|y) := \mathbb{E}(X|Y = y)$. For a policy g, a PMF π and a parameter λ, we use $\mathbb{P}_\pi^{\lambda,g}(\cdot)$ (resp. $\mathbb{E}_\pi^{\lambda,g}(\cdot)$) to indicate that the probability (resp. expectation) depends on the choice of g, π and the parameter λ. We use $\mathbf{1}_{\{x\}}(y)$ to denote the indicator that $X = x$ is in the event $\{Y = y\}$.

CHAPTER II

Centralized Stochastic Control: Multi-State Channel Sensing

2.1 Introduction

2.1.1 Motivation

Consider a communication system, shown in Fig. 2.1, consisting of N independent channels. Each channel is modeled as a K-state (K finite) Markov chain (M.C.) with known matrix of transition probabilities. At each time period a user selects one channel to sense and uses it to transmit information. A reward depending on the state of the selected channel is obtained for each transmission. The objective is to design a channel sensing policy that maximizes the expected total reward (respectively, the expected total discounted reward) collected over a finite (respectively, infinite) time horizon.

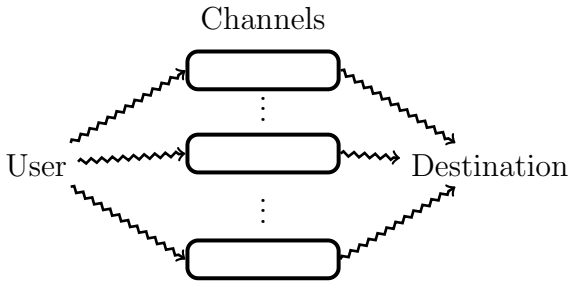

Figure 2.1: The channel sensing problem.

The above channel sensing problem arises in cognitive radio networks, opportunistic scheduling over fading channels, as well as on resource-constrained jamming ([11]). In cognitive radio networks a secondary user may transmit over a channel only when the channel is not occupied by the primary user. Thus, at any time instant t, state 1 of the M.C. describing the channel can indicate that the channel is occupied at t by the primary user, and states 2 through K indicate the quality of the channel that is available to the secondary user at t. In opportunistic transmission over fading channels, states 1 through K of the M.C. describe, at any time instant, the quality of the fading channel. In resource-constrained jamming a jammer can only jam one channel at a time, and any given jamming/channel sensing policy results in an expected reward for the jammer due to successful jamming. The physical channels in all of the above problems have memory. Introducing a finite state (K-state) Markovian model for each channel allows us to capture the effect of the channel's memory on its current quality by allowing K to take large values[1].

This channel sensing problem is also an instance of restless bandit problems ([28, 29]). Restless bandit problems arise in many areas, including wired and wireless communication systems, manufacturing systems, economic systems, statistics, biomedical engineering, business, computer science, information systems etc. (see [28, 29]).

[1]We can create a Markovian model of a finite-memory system by appropriate state expansion.

The problem described above can be formulated as a Partially Observed Markov Decision Process (POMDP) (see [30]) and can be solved, for any selection of the channels' transition probabilities and any selection of the reward process, by numerical methods. Such an approach has two drawbacks: (i) it does not provide any insight into the nature of optimal sensing strategies; (ii) it has very high computational complexity (PSPACE-complete, see [10]). For this reason we focus on identifying instances of the general problem where it is possible to explicitly characterize optimal sensing strategies. In this chapter we discover sets of conditions under which the optimal sensing strategy is the myopic policy, that is, the policy that selects at every time instant the best (in the sense of stochastic order [31]) channel. We also develop a MDP approximation for the channel sensing problem resulting in a dynamic programming algorithm that is computational feasible for a small number of channels.

2.1.2 Related Work

The channel sensing problem has been studied in [30] using a POMDP framework. For channels described by two-state Markov chains (henceforth called two-state channels), the myopic policy was studied in [32], where its optimality was established when the number of channels is two. For more than two channels, the optimality of the myopic policy was proved in [33] under certain conditions on channel parameters. This result for two-state channels was extended in [27] under a relaxed "positively correlated" condition. In [34], under the same "positively correlated" channel condition, the myopic policy was proved to be optimal for two-state channels when the user can select multiple channels at each time instance.

For general restless bandit problems, there is a rich literature; however, contrary to classical multi-armed bandit problems (see [29] and [35]), the structure (if any) of optimal strategies for general restless bandit problems is not currently known. To gain insight into the nature of restless bandit problems, research has focused on identifying

instances where an optimal strategy or qualitative properties of optimal strategies can be explicitly determined. In [28] it has been shown that the Gittins index rule (see [29] and [35] for the definition of the Gittins index rule) is not optimal for general restless bandit problems. Moreover, this class of problems is PSPACE-hard in general [10]. In [28] Whittle introduced an index policy (referred to as Whittle's index) and an "indexability condition"; the asymptotic optimality of Whittle's index was addressed in [36]. Issues related to Whittle's indexability condition were discussed in [28, 29, 36–39]. For the two-state channel sensing problem, Whittle's index was computed in closed-form in [38, 39], where performance simulation of that index was provided. For some special classes of restless bandit problems, the optimality of index-type policies was established under certain conditions (see [40, 41]). Approximation algorithms for the computation of optimal policies for a class of restless bandit problems similar to the one studied in this chapter were investigated in [42].

2.1.3 Organization

The rest of the chapter is organized as follows. In Section 2.2, we present the model and the formulation of the optimization problem associated with the channel sensing problem. In Section 2.3, we consider the finite horizon problem and identify sets of conditions sufficient to guarantee the optimality of the myopic policy; we briefly discuss the extension of our results to the infinite horizon. In Section 2.5 we show that under one particular set of conditions the myopic policy coincides with the Gittins index rule. In Section 2.6 we develop a MDP approximation that leads to a near-optimal policy, and we compare it with the myopic policy using numerical simulations. In Section 2.7 we consider the multiple selection version of the channel sensing problem, and use a MDP approximation method to numerically find a near-optimal policy when multiple selections are available. We conclude in Section 2.8. The proofs of several technical results of this chapter appear in Appendix A.

2.2 Model and the Optimization Problem

2.2.1 The Model

Consider a communication system consisting of N identical channels as shown in Fig. 2.1. Each channel is modeled as a K-state (K finite) Markov chain (M.C.) with state space $S := \{1, 2, \ldots, K\}$ and (the same) matrix of transition probabilities P,

$$P = \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1K} \\ p_{21} & p_{22} & \cdots & p_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ p_{K1} & p_{K2} & \cdots & p_{KK} \end{bmatrix} = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_K \end{bmatrix}, \tag{2.1}$$

where $P_1, P_2, ..., P_K$ are row vectors. As pointed out in Section 2.1, channels that have memory can still be modeled by Markov chain by expanding the number of states in the M.C. to account for the channel's memory. The K-state M.C. model here captures the memory characteristics of a larger class of communication channels. We assume that the channel's quality increases as the number of its state increases. We want to use this communication system to transmit information. For that matter, at each time $t = 0, 1, ..., T$, we can select one channel, observe its state, and use it to transmit information.

Let X_t^n denote the state of channel n at time t, and let U_t denote the decision made at time t; $U_t \in \{1, 2, ..., N\}$, where $U_t = n$ means that channel n is chosen for data transmission at time t.

Initially, before any channel selection is made, we assume that we have probabilistic information about the state of each of the N channels. Specifically, we assume that at $t = 0$ the user/decision-maker knows the probability mass function (PMF) on the state space of S each of the N channels; that is, the decision-maker knows

$\pi_0 := (\pi_0^1, \pi_0^2, \ldots, \pi_0^N)$, where

$$\pi_0^n := (\pi_0^n(1), \pi_0^n(2), \ldots, \pi_0^n(K)) \in \Delta(S), \quad n = 1, 2, \ldots, N, \tag{2.2}$$

$$\pi_0^n(i) := P(X_0^n = i), \quad i = 1, 2, \ldots, K. \tag{2.3}$$

Then, in general,

$$U_0 = g_0(\pi_0), \tag{2.4}$$

$$U_t = g_t(Y^{t-1}, U^{t-1}, \pi_0), \quad t = 1, 2, \ldots, \tag{2.5}$$

$$\text{where } Y^{t-1} := (Y_0, Y_1, \ldots, Y_{t-1}), \quad U^{t-1} := (U_0, U_1, \ldots, U_{t-1}), \tag{2.6}$$

and $Y_t = X_t^{U_t}$ denotes the observation at time t; Y_t gives the state of the channel that is chosen at time t (that is, if $U_t = 2$, Y_t gives the state of channel 2 at time t).
Let $R(t)$ denote the reward obtained by the transmission at time t. We assume that $R(t)$ depends on the state of the channel chosen at time t. That is

$$R(t) = R_i, \quad i = 1, 2, \ldots, K, \tag{2.7}$$

if the state of the channel chosen at t is i. Let $R := [R_1, R_2, R_3, \ldots, R_K]^T$. Since channel's quality increases as the number of its state increases, we have $R_1 \leq R_2 \leq \cdots \leq R_K$.

2.2.2 The Optimization Problem

Under the above assumptions, the objective is to solve (i) the finite horizon (T) optimization problem (P1):

Problem (P1)

$$\max_{g \in \mathcal{G}_s} \mathbb{E}^g \left[\sum_{t=0}^{T} \beta^t R(t) \right], \tag{2.8}$$

and (ii) its infinite horizon counterpart, problem (P2):

Problem (P2)

$$\max_{g \in \mathcal{G}_s} \mathbb{E}^g \left[\sum_{t=0}^{\infty} \beta^t R(t) \right], \tag{2.9}$$

where β is the discount factor ($0 < \beta \leq 1$) and $\mathcal{G}_s$ is the set of separated policies $g := (g_0, g_1, \dots)$ (see [1], Chapter 6), that are such that

$$U_t = g_t(\pi_t) \text{ for all } t, \tag{2.10}$$

$$\pi_t := (\pi_t^1, \pi_t^2, \dots, \pi_t^N) \in \Delta(S), \tag{2.11}$$

$$\pi_t^n := (\pi_t^n(1), \pi_t^n(2), \dots, \pi_t^n(K)), \quad n = 1, 2, \dots, N, \tag{2.12}$$

$$\pi_t^n(i) := P(X_t^n = i | Y^{t-1}, U^{t-1}), \quad i = 1, 2, \dots, K, \tag{2.13}$$

and π_t evolves as follows. If $U_t = n, Y^n = i$, then

$$\pi_{t+1}^n = P_i, \tag{2.14}$$

$$\pi_{t+1}^j = \pi_t^j P, \quad \text{for all } j \neq n. \tag{2.15}$$

2.2.3 Characteristics of the Optimization Problem

The optimization problems (P1) and (P2) formulated above are POMDPs; they can be solved by numerical methods, but such an approach has the drawbacks pointed out in Section 2.1.1.

Problems (P1) and (P2) can also be viewed as an instance of restless bandit

problems as follows. We can view the N channels as N arms with their PMFs as the states of the arms. The decision maker knows perfectly the states of the N arms at every time instant. One arm is operated (selected) at each time t, and an expected reward depending on the state of the selected arm is received. If arm n (channel n) is not selected at t, its PMF π_t^n evolves according to

$$\pi_{t+1}^n = \pi_t^n P; \tag{2.16}$$

if arm n (channel n) is selected at t, its PMF evolves according to

$$\pi_{t+1}^n = P_{Y_t}, \quad \mathbb{P}(Y_t = x) = \pi_t^n(x). \tag{2.17}$$

Since the selected bandit process evolves in a way that differs from the evolution of the non-selected bandit processes, this problem is a restless bandit problem.

In general, restless bandit problems are difficult to solve because forward induction (the solution methodology for the classical multi-armed bandit problem) does not result in an optimal policy [29]. Consequently, optimal policies may not be of the index type, and the form of optimal policies for general restless bandit problems (hence, the channel sensing problem) is still unknown.

To gain insight into the nature of the channel sensing problem (as well as general restless bandit problems), it is important to discover special instances of the problem where it is possible to explicitly determine optimal strategies or the structure of optimal strategies. For this season, in this chapter we focus on the "myopic policy" and we discover sets of conditions under which it is optimal. We define the myopic policy as follows. We define the concept of stochastic dominance/order (see [31]). Stochastic dominance $\geq_{st}$ between two row vectors $x, y \in \Delta(S)$ is defined as follows:

$x \geq_{st} y$ if

$$\sum_{j=i}^{K} x(j) \geq \sum_{j=i}^{K} y(j), \quad \text{for } i = 2, 3, \ldots, K. \tag{2.18}$$

Definition II.1. The myopic policy $g^m := (g_0^m, g_1^m, \ldots, g_T^m)$ is the policy that selects at each time instant the channel with the largest expected reward; that is

$$g_t^m(\pi_t) = i \qquad \text{if } \pi_t^i R \geq \pi_t^j R \quad \forall j \neq i. \tag{2.19}$$

Equivalently, from properties of stochastic order (see [31]), the myopic policy selects at each time instant the best (in the sense of stochastic order) channel; that is,

$$g_t^m(\pi_t) = i \qquad \text{if } \pi_t^i \geq_{st} \pi_t^j \quad \forall j \neq i. \tag{2.20}$$

2.3 Analysis of the Finite Horizon Problem

We will prove the optimality of the myopic policy g^m for Problem (P1) under certain specific assumptions on the structure of the Markov chains describing the channels, on the instantaneous rewards $R = [R_1, R_2, R_3, \ldots, R_K]^T$ and on the initial PMFs $\pi_0^1, \pi_0^2, \ldots, \pi_0^N$. We proceed as follows. In Section 2.3.1 we discuss why the problem under consideration is not a trivial extension of the instance where each channel has only two states (studied in [27]). This discussion helps to justify the key assumptions/conditions we make in Section 2.3.2. These assumptions/conditions reduce to those of [27] when $K = 2$. The main result of the chapter is stated in Section 2.3.3; its proof appears in Section 2.3.4 to 2.3.7. The key features of the solution approach and the role of the conditions in the approach are discussed in Section 2.3.8.

2.3.1 Difficulties in Establishing the Optimality of the Myopic Policy

The situation where each channel has two states, i.e. $K = 2$, has been previously investigated in [27] where the optimality of the myopic policy is established under some conditions. In the two-state channels situation, the PMF in equation (2.12) (called the information state of the POMDP, see [1]) can be described by a number, the conditional probability of the "best state". As a result of this feature, the information states of all channels can be totally ordered at any time regardless of channels' evolution. Such an ordering is needed for the derivation of the results in [27]. In our problem the information state defined by equation (2.12) is a $(K-1)$-dimensional vector; $(K-1)$-dimensional vectors can not, in general, be ordered at every time instant. This difference between the information state of two-state channels and the one in this chapter results in a lot of complications in extending the results of [27] to multi-state channels.

In general, an extension of the results on the optimality of the myopic policy for two-state channels to multi-state channels would require: (i) An ordering of the channels' information states (PMFs defined by eq. (2.12)) at every time instant. Such an ordering can only be ensured under certain conditions (Conditions (A1)-(A3) appearing in Section 2.3.2) on the evolution of the channels. (ii) If the myopic policy is to be optimal, the instantaneous expected gain incurred by choosing the best channel (say channel n) versus any other channel (say channel m) must overcompensate expected future losses in performance resulting in when channel m is chosen instead of channel n. We have K channel states and this leads to $K-1$ inequalities in Condition (A4) (appearing in Section 2.3.2) on the separation of instantaneous rewards. Condition (A4) describes how much the instantaneous rewards obtained in states i and $i-1, i = 2, 3, ..., K$, should be separated so as to ensure the optimality of the myopic policy.

The above discussion provides the rationale for Conditions (A1)-(A4) appearing

below.

2.3.2 Key Assumptions/Conditions

We make the following assumptions/conditions

(A1)

$$P_K \geq_{st} P_{K-1} \geq_{st} \cdots \geq_{st} P_1. \tag{2.21}$$

Note that the quality of a channel state increases as its number increases. Assumption (A1) ensures that the higher the quality of the channel's current state the higher is the likelihood that the next channel state will be of high quality. This requirement is the same as the "positively correlated" condition when $K = 2$ in [27].

(A2) Let $\Delta(S)P := \{\pi P : \pi \in \Delta(S)\}$. At time 0,

$$\pi_0^1, \pi_0^2, \ldots, \pi_0^N \in \Delta(S)P, \tag{2.22}$$

$$\text{and } \pi_0^1 \leq_{st} \pi_0^2 \leq_{st} \cdots \leq_{st} \pi_0^N. \tag{2.23}$$

Assumption (A2) states that initially the channels can be ordered in terms of their quality, expressed by the PMF on S. Moreover, the initial PMFs of the channels are in $\Delta(S)P$. The requirement expressed by (2.22) is always satisfied since the channels evolve before we begin sensing them. Requirement (2.22) also ensures that the initial PMFs on the channel states are in the same space as all subsequent PMFs.

(A3) There exists some L, $2 \leq L \leq K$ such that

$$P_1 P \geq_{st} P_{L-1}, \tag{2.24}$$

$$P_K P \leq_{st} P_L. \tag{2.25}$$

Assumption (A3) along with (A2) ensure that, any PMF π reachable from a non-selected channel has quality between P_{L-1} and P_L, that is $P_L \geq_{st} \pi \geq_{st} P_{L-1}$ (see also Property II.4, Section 2.3.4).

As pointed out in Section 2.3.1, (A1)-(A3) ensure that the channels' information states are ordered at any time t (see Property II.5, Section 2.3.4).

(A4)

$$\begin{aligned} R_i - R_{i-1} \geq & \beta(P_i - P_{i-1})M \\ \geq & \beta(P_i - P_{i-1})U \geq 0, \quad \text{for } i \neq L, \end{aligned} \tag{2.26}$$

$$R_L - R_{L-1} \geq \beta(h - P_{L-1}R) \geq 0, \tag{2.27}$$

where M and U are vectors given by

$$M := U + \beta \sum_{i \geq L} p_{Ki} P U, \tag{2.28}$$

$$U_i := R_i \quad \text{for } i = 1, 2, \ldots, L-1, \tag{2.29}$$

$$U_i := R_i + \beta(P_i - P_{L-1})U, \quad \text{for } i = L, L+1, \ldots, K, \tag{2.30}$$

and h is given by

$$h = \frac{P_K R - \beta \sum_{i<L} p_{Ki} P_i R}{1 - \beta \sum_{i<L} p_{Ki}}. \tag{2.31}$$

Assumption (A4) states that the instantaneous rewards obtained at different

states of the channel are sufficiently separated (see (2.26) and (2.27)). The reason for such a separation was discussed in Section 2.3.1.

We note that (A1)-(A4) describe sets of sets of assumptions/conditions; for every value of $L, L = 2, 3, ..., K$, we have a distinct set of conditions.

Before we proceed with the analysis of Problem (P1) based on conditions (A1)-(A4), we show that (A1)-(A4) can be simultaneously satisfied. Consider the following situation:

$$K = 5, L = 5, N = 6, \beta = 1 \tag{2.32}$$

$$P = \begin{bmatrix} P_1 \\ P_2 \\ \vdots \\ P_5 \end{bmatrix} = \begin{bmatrix} 0.0656 & 0.0458 & 0.1044 & 0.4745 & 0.3096 \\ 0.0655 & 0.0458 & 0.1030 & 0.4454 & 0.3403 \\ 0.0652 & 0.0457 & 0.0966 & 0.4019 & 0.3907 \\ 0.0434 & 0.0336 & 0.1126 & 0.4102 & 0.4001 \\ 0.0206 & 0.0205 & 0.0142 & 0.4475 & 0.4972 \end{bmatrix}, \tag{2.33}$$

with

$$R = \begin{bmatrix} 0 & 1 & 2 & 3 & 4 \end{bmatrix}^T \tag{2.34}$$

$$\pi_0^1 = \pi_0^2 = P_1, \pi_0^3 = P_2, \pi_0^4 = P_3, \pi_0^5 = P_4, \pi_0^6 = P_5 \tag{2.35}$$

By their definition, $P_1, P_2, ..., P_5$ satisfy (A1). By the definition of $\pi_0^1, \pi_0^2, ..., \pi_0^6$ and the definition of $\Delta(S)P$, (A2) is satisfied.

By direct computation we can show that

$$\begin{aligned} P_1 P &= \begin{bmatrix} 0.0411 & 0.0322 & 0.0795 & 0.4267 & 0.4205 \end{bmatrix} \\ &\geq_{st} \begin{bmatrix} 0.0434 & 0.0336 & 0.1126 & 0.4102 & 0.4001 \end{bmatrix} = P_4, \end{aligned} \tag{2.36}$$

Moreover, $P_5P = p_{51}P_1 + p_{52}P_2 + +... + p_{55}P_5 \leq_{st} P_5$. Therefore, (A3) is satisfied. By direct computation, we get

$$U = \begin{bmatrix} 0 & 1 & 2 & 3 & 4.3214 \end{bmatrix}^T \tag{2.37}$$

$$M = \begin{bmatrix} 1.4997 & 2.5206 & 3.5577 & 4.6003 & 6.0815 \end{bmatrix}^T \tag{2.38}$$

$$h = 3.7776, \tag{2.39}$$

So we can compute

$$\beta(P_2 - P_1)M = 0.0470 \leq R_2 - R_1 \tag{2.40}$$

$$\beta(P_3 - P_2)M = 0.0829 \leq R_3 - R_2 \tag{2.41}$$

$$\beta(P_4 - P_3)M = 0.0897 \leq R_4 - R_3 \tag{2.42}$$

$$\beta(h - P_4R) = 0.7766 \leq R_5 - R_4 \tag{2.43}$$

Therefore, (A4) is satisfied.

Assumptions (A1)-(A4) are also satisfied when $R, P, \pi_0^1, \pi_0^2, ..., \pi_0^6$, chosen as above, are slightly perturbed. It is also possible to find other ranges of values of $R, P, \pi_0^1, \pi_0^2, ..., \pi_0^6$ which satisfy (A1)-(A4).

When $K = 2$ the above Conditions (A1)-(A4) reduce to those of [27] as follows. When $K = 2, L = K$. Then, our Conditions (A1)-(A4) reduce to

$$p_{2,2} \geq p_{1,2}, \tag{2.44}$$

$$\pi_0^n = (1 - p^n, p^n), \quad p_{1,2} \leq p^n \leq p_{2,2} \text{ for } n = 1, 2, \ldots, N, \tag{2.45}$$

$$p^1 \leq p^2 \leq \ldots \leq p^N. \tag{2.46}$$

Condition (2.44) is precisely the "positively correlated" condition in [27]. Condition (2.45) is satisfied, if the channels evolve before we begin sensing them (before time

$t = 0$). Condition (2.46) is always satisfied by renumbering of the channels.

2.3.3 The Main Result

The main result we establish in this chapter is given by Theorem II.2 below.

Theorem II.2. *Under assumptions (A1)-(A4), the myopic policy g^m, that is, the policy that selects at every time instant the best (in the sense of stochastic order) channel is optimal for Problem (P1).*

We proceed to establish the optimality of the myopic policy g^m as follows. In sections 2.3.4-2.3.6 we develop preliminary results needed for the proof of Theorem II.2. Specifically: In section 2.3.4 we present three properties of the evolution of the PMFs on the channel states. In section 2.3.5 we present a property of the instantaneous expected reward. In section 2.3.6 we define a class of ordering-based channel sensing policies $\mathcal{G}^O$ which includes the myopic policy g^m; using the results of sections 2.3.4 and 2.3.5 we discover four properties of the expected reward resulting from any policy in $\mathcal{G}^O$. In section 2.3.7 we use the results of section 2.3.6 to establish the optimality of the myopic policy for Problem (P1). The properties' proofs appear in Appendix A.

2.3.4 Properties of the Channels' Evolution

Under assumptions/conditions (A1)-(A4) stated in section 2.3.2, the following properties hold.

Proposition II.3. *Let $x, y \in \Delta(S)$. Under Assumption (A1),*

$$x \geq_{st} y \Longrightarrow xP \geq_{st} yP. \tag{2.47}$$

An implication of Property II.3 is the following. If at any time t the information states of two channels (expressed by the PMFs on their state space) are stochastically

ordered and none of these channels is sensed at t, then the same stochastic order between the information states at time $t+1$ is maintained.

Proposition II.4. *Let* $\pi = xP^2 \in \Delta(S)P^2$, $\Delta(S)P^2 := \{\pi = xP^2, x \in \Delta(S)\}$. *Under (A1)-(A3),*

$$P_L \geq_{st} xP^2 \geq_{st} P_{L-1}. \tag{2.48}$$

Property II.4 says the following. By Condition (A2) a channel's information state (the PMF on its state space) is always in $\Delta(S)P$. If the channel is not sensed at time t, then at time $t+1$ its information state is in $\Delta(S)P^2$, moreover it is stochastically always between P_{L-1} and P_L. If the channel is sensed at time t and its observed state is larger than or equal to L (respectively smaller than L), then at time $t+1$ this channel is in the stochastically largest (respectively stochastically smallest) information state among all channels.

Proposition II.5. *Under (A1)-(A3), we have either* $\pi_t^n \leq_{st} \pi_t^m$ *or* $\pi_t^m \leq_{st} \pi_t^n$ *for all* $n, m \in \{1, 2, \ldots, N\}$ *for all* t.

Property II.5 states that under (A1)-(A3) the information states of all channels can be ordered stochastically at all times.

2.3.5 A Property of the Instantaneous Expected Reward

A direct consequence of Condition (A4) is the following property of the instantaneous expected reward:

Proposition II.6. *Let* $x, y \in \Delta(S)$. *Let* v *be a column vector in increasing order, i.e.* $v_i \geq v_{i-1}$ *for* $i = 2, 3, \ldots, K$. *If* $x \geq_{st} y$, *we have*

(i) $(x - y)v \geq 0$.

(ii) $(x-y)M \geq (x-y)U \geq (x-y)R \geq 0$, *where* M, U, R *are defined by equations (2.26)-(2.30).*

(iii) $(x-y)M \geq \beta(x-y)PM$.

(iv) If $x(i) = y(i)$ *for all* $i \geq L$ *or* $x(i) = y(i)$ *for all* $i < L$, *then*

$$(x-y)R \geq \beta(x-y)PM. \tag{2.49}$$

Part (i) of Property II.6 says the following. Consider a reward vector such that the reward increases as the quality of the channel state increases. Then the expected reward increases as the information state of the channel increases stochastically.

Part (ii) is a restatement of part (i) when the reward vector v takes the values $M-U, U-R, R$.

Part (iii) can be interpreted as follows. Consider the reward vector M defined by (2.28). Consider two channels, channel i and channel j, that have information states x and y respectively, such that $x \geq_{st} y$. Consider the following scenarios: (SC1) Sense channel i first, then sense channel j; (SC2) Sense channel j first, then sense channel i. Afterwards, continue with the same channel selection sequence under both scenarios. Then part (iii) of Property II.6 asserts that scenario (SC1) is better than scenario (SC2) in terms of the expected accumulated rewards; that is, it is better to sense the best (in the sense of stochastic order) channel first.

Part (iv) has an interpretation similar to that of part (iii) when x, y satisfy the condition of part (iv).

2.3.6 Properties of the Reward Associated with Ordering-based Channel Sensing Polices

In this section we introduce ordering-based policies and study their properties. The reason for considering this class of policies is because under Conditions (A1)-

(A4) we obtain the following: (i) The performance of any sensing policy can be upper-bounded by an appropriately chosen ordering-based policy (see Section 2.3.7); thus, for the solution of the original optimization problem (Problem (P1)) we can restrict attention to ordering-based policies. (ii) The myopic policy is an optimal ordering-based policy. Combining (i) and (ii) we establish the optimality of the myopic policy for Problem (P1).

We note that Properties 1-4, developed so far, are essential for the discovery of the properties of ordering-based policies that lead eventually to the solution of Problem (P1) (see discussion in Section 2.3.8).

Let $\mathcal{O}$ be the set of all orderings/permutations of the N channels $\{1, 2, \ldots, N\}$. Consider the ordering-based selection function $\hat{g} : \mathcal{O} \mapsto \{1, 2, \ldots, N\}$ and the ordering update mapping $\hat{m} : \mathcal{O} \times \{1, 2, \ldots, K\} \mapsto \mathcal{O}$ defined as follows. For every $O := (O(1), O(2), \ldots, O(N)) \in \mathcal{O}$,

$$\hat{g}(O) = O(N), \tag{2.50}$$

$$\hat{m}(O, y) = \begin{cases} O & \text{if } y \geq L, \\ SO & \text{if } y < L, \end{cases} \tag{2.51}$$

where S is the cyclic shift operator on $\mathcal{O}$ such that

$$SO =: (O(N), O(1), O(2), \ldots, O(N-1)). \tag{2.52}$$

Given a channel ordering $O_t \in \mathcal{O}$ at time t, we define an ordering-based channel

sensing policy $g_{t:T}^{O_t} := (g_t^{O_t}, g_{t+1}^{O_t}, \ldots, g_T^{O_t})$ as follows.

$$U_t = g_t^{O_t}(O_t) = \hat{g}(O_t) = O(N), \tag{2.53}$$

$$O_s = \hat{m}(O_{s-1}, Y_{s-1}), \quad \text{for } s = t+1, t+2, \ldots, T, \tag{2.54}$$

$$\begin{aligned} U_s =& g_s^{O_t}(Y_{t:s-1}, U_{t:s-1}) \\ =& g_s^{O_t}(O_s) = \hat{g}(O_s), \quad \text{for } s = t+1, t+2, \ldots, T. \end{aligned} \tag{2.55}$$

At time $s, t \le s \le T$, $g_s^{O_t}$ chooses the last channel in O_s; the ordering O_s is shifted to the right by the update mapping $\hat{m}$ whenever the observed state is less than L, and remains the same otherwise. As a result of the above specification of $g_{t:T}^{O_t}$, if at time t channel n is on the right of channel m in the ordering O_t, channel n will be sensed by policy $g_{t:T}^{O_t}$ before channel m.

Note that, the policy $g_{t:T}^{O_t}$ is not a separated policy in general. However, if the ordering $O_0 = (O_0(1), O_0(2), \ldots, O_0(N))$ at time 0 is such that $\pi_0^{O_0(1)} \le_{st} \pi_0^{O_0(2)} \le_{st} \cdots \le_{st} \pi_0^{O_0(N)}$, then $g_{0:T}^{O_0}$ is the myopic policy g^m, therefore; $g_{0:T}^{O_0} = g^m \in \mathcal{G}_s$, as the following property shows.

Proposition II.7. *At time $t = 0$ consider the ordering O_0 such that $\pi_0^{O_0(1)} \le_{st} \pi_0^{O_0(2)} \le_{st} \cdots \le_{st} \pi_0^{O_0(N)}$. Then, the ordering based policy $g_{0:T}^{O_0}$ is just the myopic policy g^m.*

The validity of Property II.7 crucially depends on Properties II.3 and II.4, which say that stochastic order is maintained under the evolution of unobserved channels (Property II.3), and the observed channel is either the stochastically best or the stochastically worst among all channels (Property II.4). Without Properties II.3 and II.4 the myopic policy is not an ordering-based policy.

Define $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ to be the expected reward collected from time t up

to and including T due to the ordering-based policy $g_{t:T}^{O_t}$. That is,

$$V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) := \mathbb{E}^{g_{t:T}^{O_t}} \left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N \right]. \tag{2.56}$$

Then, $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ can be written recursively as follows.

$$V_T(O_t, \pi_T^1, \pi_T^2, \ldots, \pi_T^N) = \pi_T^{O_t(N)} R, \tag{2.57}$$

$$\begin{aligned} & V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ =& \pi_t^{O_t(N)} R + \beta \sum_{i<L} \pi_t^{O_t(N)}(i) V_{t+1}(SO_t, \pi_{t+1}^1, \ldots, \pi_{t+1}^N) \\ & + \beta \sum_{i \geq L} \pi_t^{O_t(N)}(i) V_{t+1}(O_t, \pi_{t+1}^1, \ldots, \pi_{t+1}^N), \end{aligned} \tag{2.58}$$

$$\text{where } \pi_{t+1}^n = \begin{cases} P_i & \text{for } n = O_t(N), \\ \pi_t^n P & \text{otherwise.} \end{cases} \tag{2.59}$$

The function $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ defined above possesses Properties II.8-II.11 below. We will explain the role of these properties in Section 2.3.8 after we prove the main result on the optimality of the myopic policy in Section 2.3.7.

Proposition II.8. *Let $\hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N \in \Delta(S)P$ and $O_t \in \mathcal{O}$.*
Define

$$L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) := V_t(O_t, \hat{\pi}_t^1, \pi_t^2, \ldots, \pi_t^N) - V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N). \tag{2.60}$$

If $\hat{\pi}_t^1 \geq_{st} \pi_t^1$, and $O_t(n) = 1$, then for all $m < n$

$$\begin{aligned} 0 \leq & L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) - L_t(S^{-m} O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ \leq & (\hat{\pi}_t^1 - \pi_t^1) U, \end{aligned} \tag{2.61}$$

where $S^{-m}O_t$ is the counter-clockwise cyclic shift of O_t by m positions, that is,

$$S^{-m}O_t = (O_t(m+1), O_t(m+2), \ldots, O_t(N), O_t(1), \ldots, O_t(m)). \tag{2.62}$$

Proposition II.9. *For $O_t \in \mathcal{O}$, define the operator W_{nm} as follows.*

$$W_{nm}O_t(i) := \begin{cases} O_t(n) & \text{for } i = m, \\ O_t(m) & \text{for } i = n, \\ O_t(i) & \text{otherwise.} \end{cases} \tag{2.63}$$

If $\hat{\pi}_t^1 \geq_{st} \pi_t^1$, and $O_t(n) = 1$, then for $m < n$

$$\begin{aligned} 0 \leq & L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) - L_t(W_{nm}O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\ \leq & (\hat{\pi}_t^1 - \pi_t^1)M. \end{aligned} \tag{2.64}$$

The meaning of Properties II.8 and II.9 is the following. Restrict attention to ordering-based policies. Take any channel, say channel 1. Replace it with a better quality (in the sense of stochastic order) channel. Such a replacement will result in an improvement in performance. This improvement is different for different channel orderings. The earlier channel 1 is used (that is, the closer to the right-most position in the ordering channel 1 is) the higher is the improvement. Properties II.8 and II.9 also provide bounds on the difference between maximum and minimum improvement. These bounds are useful in proving Properties II.8 and II.9 by induction.

Proposition II.10. *If $\pi_t^{O_t(n)} \geq_{st} \pi_t^{O_t(m)}$, then for $m < n$ then*

$$V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \geq V_t(W_{nm}O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N). \tag{2.65}$$

Property II.10 states that if the position of two channels in any arbitrary but fixed channel ordering are interchanged so that the better (in the stochastic order sense)

channel comes closer to the right-most position (i.e. it is used earlier) in the new ordering, the performance due to the ordering-based policy improves.

Proposition II.11. *For $O_t \in \mathcal{O}$, define the operator A_{nm} as follows.*

$$A_{nm}O_t(i) := \begin{cases} O_t(n) & \text{for } i = m, \\ O_t(i-1) & \text{for } i = m+1, m+2, \ldots, n, \\ O_t(i) & \text{otherwise.} \end{cases} \tag{2.66}$$

If $\pi_t^1 \leq_{st} \pi_t^1 P$, and $O_t(n) = 1$, then

$$V_t(A_{nm}O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) - V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \leq h - \pi_t^1 P^{N-n} R. \tag{2.67}$$

Property II.11 states the following. Suppose that a channel, say channel 1, is such that as long as it is not sensed its quality is continuously improving (i.e. its PMF is continuously increasing stochastically). Then, no matter how late this channel is sensed (that is, no matter how much we move the channel to the left from its initial position in the original channel ordering) the change in performance due to an ordering-based policy can not exceed a certain bound, given by the right hand side of (2.67).

Properties II.8-II.11 are proved simultaneously in Appendix A. The idea of their proof may be useful in stochastic scheduling problems where the optimality of "list polices" ([43]) is investigated. In the analysis of "list polices", it is important to compare the performance due to different orders of task processing/scheduling. To do this we consider an initial ordering of the tasks to be processed. We perturb the ordering and study the resulting change in performance. Several types of perturbation need to be examined. Typical types of such perturbations are described in the statements of Properties II.8-II.11. The proof of Properties II.8-II.11 indicates that such perturbations can not be analyzed in isolation but have to be considered

simultaneously.

2.3.7 Proof of the Main Result (Theorem II.2)

Proof. We proceed by induction.

At T, the expected reward is the instantaneous expected reward. Since by part (ii) of Property II.6 a better channel (in the sense of stochastic order) gives larger instantaneous expected reward, the myopic policy g^m is optimal at T. This establishes the basis of induction.

The induction hypothesis is that the myopic policy g^m is optimal at $t+1, t+1, \ldots, T$.

Without loss of generality, we assume $\pi_t^1 \leq_{st} \pi_t^2 \leq_{st} \cdots \leq_{st} \pi_t^N$. Consider any policy g. If g picks channel n at time t, then the expected reward collected from t on due to the policy g is given by

$$\begin{aligned}
&\mathbb{E}^g\left[\sum_{l=t}^{T}\beta^{l-t}R(l)|\pi_t^1,\ldots,\pi_t^N\right]\\
=&\pi^n R+\sum_{i=1}^{K}\pi_t^n(i)\mathbb{E}^g\left[\sum_{l=t+1}^{T}\beta^{l-t}R(l)|\pi_{t+1}^n=P_i,\pi_{t+1}^m=\pi_t^m P \text{ for } m\neq n\right]\\
\leq&\pi^n R+\sum_{i=1}^{K}\pi_t^n(i)\mathbb{E}^{g^m}\left[\sum_{l=t+1}^{T}\beta^{l-t}R(l)|\pi_{t+1}^n=P_i,\pi_{t+1}^m=\pi_t^m P \text{ for } m\neq n\right]\\
=&\pi_t^n R+\beta\sum_{i<L}\pi_t^n(i)V_{t+1}(SO_t,\pi_{t+1}^1,\ldots,\pi_{t+1}^N)+\beta\sum_{i\geq L}\pi_t^n(i)V_{t+1}(O_t,\pi_{t+1}^1,\ldots,\pi_{t+1}^N)\\
=&V_t(O_t,\pi_t^1,\ldots,\pi_t^N). \qquad (2.68)
\end{aligned}$$

The inequality in (2.68) follows from the induction hypothesis and the ordering $O_t := (1,2,\ldots,n-1,n+1,\ldots,N,n)$.

Since $\pi_t^n \leq_{st} \pi_t^m$ for all $m = n+1, n+2, \ldots, N$, repeatedly applying Property

II.10 we get

$$\begin{aligned} & V_t(O_t, \pi_t^1, \ldots, \pi_t^N) \\ \leq & V_t((1, 2, \ldots, n-1, n, n+1, \ldots, N), \pi_t^1, \ldots, \pi_t^N) \\ = & \mathbb{E}^{g^m}\left[\sum_{l=t}^{T} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right]. \end{aligned} \tag{2.69}$$

Combining (2.68) (2.69) we obtain

$$\mathbb{E}^{g}\left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right] \leq \mathbb{E}^{g^m}\left[\sum_{l=t}^{T} \beta^{l-t} R(l) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N\right], \tag{2.70}$$

which completes the proof. □

2.3.8 Discussion

The key steps in establishing the optimality of the myopic policy, under the assumptions made in the problem formulation, are the following:

(K1) The assertion that the performance of any separated policy can be upper-bounded by the performance of an ordering-based policy. Consequently, for the solution of the original optimization problem, one can restrict attention to ordering-based policies.

(K2) The assertion that the performance of an ordering-based policy improves when a better (in the sense of stochastic order) channel is used earlier. This assertion implies the optimality of the myopic policy.

The assertion of (K1) is established in Theorem II.2 (its induction step). The assertion of (K2) is established by Property II.10, provided that the myopic policy is an ordering-based policy, and that stochastic order is maintained among all channels at every time. The fact that the myopic policy is an ordering-based policy is ensured by

Property II.7. The existence of a stochastic ordering among all channels at any time t is ensured by Property II.5. To establish these properties we need Properties 1-9.

We now elaborate on the interdependence of Properties 1-9. Property II.5, which asserts that channels can be ordered stochastically, is a consequence of Properties II.3 and II.4 for the unobserved channels and the observed channel, respectively. Properties II.3 and II.4 also ensure that the myopic policy g^m belongs to the class of ordering-based policies (Property II.7). Property II.10 is a special case of Property II.9 when $\hat{\pi}_t^1 = \pi_t^{O_t(m)} \geq_{st} \pi_t^1 = \pi_t^{O_t(n)}$. Property II.9 is coupled with Properties II.8 and II.11, that is, Properties II.8, II.9 and II.11 need to be proven simultaneously. The proof of Properties II.8, II.9 and II.11 requires Property II.6.

The upper bounds that appear in Properties II.8, II.9 and II.11 are essential in establishing the optimality of the myopic policy. These bounds along with Condition (A4) ensure that the instantaneous advantage in expected reward obtained by the use of the myopic policy g^m over any other policy g, overcompensates any future possible expected reward losses of g^m as compared to g.

2.4 The Infinite Horizon Problem

For the infinite horizon Problem (P2) we have the following theorem.

Theorem II.12. *Under assumptions (A1)-(A4), the myopic policy g^m is optimal for Problem (P2).*

Proof. From the theory of stochastic control [1] we know that for Problem (P2) there exists a separated stationary policy g^* that maximizes the total expected discounted reward.

Let $\pi := (\pi^1, \pi^2, ..., \pi^N)$; for any stationary separated policy g let

$$J_\beta^g(\pi) := \mathbb{E}^g \left[\sum_{t=0}^{\infty} \beta^t R(t) | \pi_0 = \pi \right]. \tag{2.71}$$

Then the dynamic program for Problem (P2) is

$$J_\beta^{g^*}(\pi) = \max_{n=1,2,\ldots,N} \left\{ \pi^n R + \beta \mathbb{E}\left[J_\beta^{g^*}(\pi_1) | \pi_0 = \pi, U_0 = n \right] \right\}, \tag{2.72}$$

where π_0, π_1 are defined by (2.11)-(2.13). The myopic policy g^m that is optimal for the finite horizon T problem (by Theorem II.2) satisfies the dynamic program

$$J_{\beta,T}^{g^m}(\pi) = \max_{n\in\{1,2,\ldots,N\}} \left\{ \pi^n R + \beta \mathbb{E}\left[J_{\beta,T-1}^{g^m}(\pi_1) | \pi_0 = \pi, U_0 = n \right] \right\}, \tag{2.73}$$

where

$$J_{\beta,T}^{g^m}(\pi) := \mathbb{E}^{g^m}\left[\sum_{t=0}^{T} \beta^t R(t) | \pi_0 = \pi \right]. \tag{2.74}$$

Since the reward $R(t) \leq R_K$ is bounded, by the bounded convergence theorem we get

$$\begin{aligned} J_\beta^{g^m}(\pi) =& \mathbb{E}^g\left[\sum_{t=0}^{\infty} \beta^t R(t) | \pi_0 = \pi \right] \\ =& \lim_{T\to\infty} \mathbb{E}^g\left[\sum_{t=0}^{T} \beta^t R(t) | \pi_0 = \pi \right] \\ =& \lim_{T\to\infty} J_{\beta,T}^{g^m}(\pi), \end{aligned} \tag{2.75}$$

Letting $T \to \infty$ in (2.73) and using the bounded convergence theorem we obtain

$$J_\beta^{g^m}(\pi) = \max_{n\in\{1,2,\ldots,N\}} \left\{ \pi^n R + \beta \mathbb{E}\left[J_\beta^{g^m}(\pi_1) | \pi_0 = \pi, U_0 = n \right] \right\}, \tag{2.76}$$

Notice that (2.76) is exactly the dynamic programming equation (2.72); therefore,

$$J_\beta^{g^m}(\pi) = J_\beta^{g^*}(\pi); \tag{2.77}$$

consequently, the myopic policy g^m is optimal for the infinite horizon problem (P2).

□

2.5 Myopic policy vs. Gittins index rule

In this section we investigate conditions under which the myopic policy coincides with the Gittins index rule.

Select a channel, say channel $n, n = 1, 2, \ldots, N$. For PMF $\pi \in \Delta(S)$, the Gittins index ([29, 35]) of channel n is defined is defined by

$$\nu^n(\pi) := \max_{\tau} \frac{\mathbb{E}^{g^\tau}\left[\sum_{t=0}^{\tau-1} \beta^t \pi_t^n R | \pi_0^n = \pi\right]}{\mathbb{E}^{g^\tau}\left[\sum_{t=0}^{\tau-1} \beta^t | \pi_0^n = \pi\right]}, \tag{2.78}$$

where τ is any stopping time with respect to $\{\pi_t^n, t = 0, 1, \ldots\}$ and g^τ chooses channel n from $t = 0$ up to $t = \tau - 1$. The Gittins index rule ([29, 35]) chooses the channel with the highest Gittins index at every time instant t.

In condition (A3) (Section 2.3.2) L is fixed; it can be any number form 2 to K. In this section we show that when $L = K$, under conditions (A1)-(A4), after time 0 the myopic policy coincides with the Gittins index rule. We establish this result via Theorems II.13 and II.14.

Theorem II.13. *(i) For $\pi \in \Delta(S)P$, $P_{K-1} \leq_{st} \pi \leq_{st} P_K$, the Gittins index $\nu(\pi)$ is given by*

$$\nu(\pi) = \frac{\pi R + \beta\pi(K)\frac{P_K R}{1-\beta p_{KK}}}{1 + \beta\pi(K)\frac{1}{1-\beta p_{KK}}}. \tag{2.79}$$

(ii) If $\pi_x, \pi_y \in \Delta(S)P$ and $P_{K-1} \leq_{st} \pi_y \leq_{st} \pi_x \leq_{st} P_K$, then $\nu(\pi_x) \geq \nu(\pi_y)$.

(iii) If $\pi \in \Delta(S)P$ and $P_{K-1} \leq_{st} \pi \leq_{st} P_K$, then $\nu(\pi) \geq \nu(P_i)$ for $i < K$.

Proof. (i) From Property II.4 and part (ii) of Property II.6 we know that

$$\pi R \leq P_K R \text{ for all } \pi \in \Delta(S)P. \tag{2.80}$$

Using (2.80) in the definition of Gittins index (2.78) we get

$$\nu(\pi) \leq P_K R \text{ for all } \pi \in \Delta(S)P. \tag{2.81}$$

Letting $\tau = 1$ in (2.78), we get an lower bound on the Gittins index of P_K

$$\nu(P_K) \geq \mathbb{E}[R(\pi_0)|\pi_0 = P_K] = P_K R. \tag{2.82}$$

Combining (2.81) and (2.82), $\nu(P_K) = P_K R$ and the PMF P_K has the largest Gittins index among all PMFs.

From Theorem 4.1 in [44] we know that the second largest Gittens index among PMFs $\{\pi, P_1, P_2, .., P_{K-1}, P_K\}$ is given by

$$\max_{x=\{\pi,P_1,P_2,..,P_{K-1}\}} \nu_K(x), \tag{2.83}$$

where

$$\nu_K(x) := \frac{A_K(x)}{B_K(x)}, \tag{2.84}$$

$$A_K(x) := xR + \beta x(K) A_K(P_K), A_K(P_K) = \frac{P_K R}{1 - \beta P_{KK}}, \tag{2.85}$$

$$B_K(x) := 1 + \beta x(K) B_K(P_K), B_K(P_K) = \frac{1}{1 - \beta P_{KK}}. \tag{2.86}$$

We now show that for $P_{K-1} \leq_{st} \pi \leq_{st} P_K$

$$\nu_K(\pi) = \max_{x=\{\pi,P_1,P_2,..,P_{K-1}\}} \nu_K(x). \tag{2.87}$$

For that matter we need to show that $\nu(\pi_x) \geq \nu(\pi_y)$ whenever $\pi_x \geq_{st} \pi_y, \pi_x, \pi_y \in \Delta(S)P$. From (2.84),

$$\begin{aligned}
\nu_K(\pi_x) =& \frac{\pi_x R + \beta \pi_x(K) A_K(P_K)}{1 + \beta \pi_x(K) B_K(P_K)} \\
=& P_K R + \frac{\pi_x R - P_K R}{1 + \beta \pi_x(K) B_K(P_K)} \\
\geq& P_K R + \frac{\pi_y R - P_K R}{1 + \beta \pi_x(K) B_K(P_K)} \\
\geq& P_K R + \frac{\pi_y R - P_K R}{1 + \beta \pi_y(K) B_K(P_K)} \\
=& \nu_K(\pi_y). \qquad (2.88)
\end{aligned}$$

The first inequality in (2.88) follows from part (ii) of Property II.6 and $\pi_x \geq_{st} \pi_y$. The second inequality in (2.88) holds because $\pi_y R - P_K R \leq 0$ as $\pi_y \leq_{st} P_K$.

Since $\pi \geq_{st} P_i$ for $i = 1, 2, \ldots, K-1$, (2.88) ensures that $\nu_K(\pi) \geq \nu_K(P_i)$ for $i = 1, 2, \ldots, K-1$. Thus, π is the PMF with the second largest Gittins index among $\{\pi, P_1, P_2, .., P_{K-1}, P_K\}$.

The Gittins index for $\pi \in \Delta(S)P, P_{K-1} \leq_{st} \pi \leq_{st} P_K$ is given by

$$\nu(\pi) = \nu_K(\pi) = \frac{\pi R + \beta \pi(K) \frac{P_K R}{1 - \beta p_{KK}}}{1 + \beta \pi(K) \frac{1}{1 - \beta p_{KK}}}. \qquad (2.89)$$

This completes the proof of (i).

(ii) If $\pi_x, \pi_y \in \Delta(S)P$ and $P_{K-1} \leq_{st} \pi_y \leq_{st} \pi_x \leq_{st} P_K$, by (2.88) and (2.89), we get

$$\nu(\pi_y) = \nu_K(\pi_y) \leq \nu_K(\pi_x) = \nu(\pi_x). \qquad (2.90)$$

(iii) From part (i) we know that for $\pi \in \Delta(S)P$ and $P_{K-1} \leq_{st} \pi \leq_{st} P_K$, π gives the second largest Gittins index among $\{\pi, P_1, P_2, .., P_{K-1}, P_K\}$. Consequently, $\nu(\pi) \geq \nu(P_i)$ for $i < K$.

□

Theorem II.14. *Under Conditions (A1)-(A4) and $L = K$, after time $t = 0$ the Gittins index rule is an optimal channel sensing policy for Problems (P1).*

Proof. Consider any time $t > 0$. If the channel observed at time $t-1$ is in state K then the PMF of that channel at t is P_K. The myopic policy senses this channel at t. The Gittins index rule senses the same channel at t as P_K is the PMF with the largest Gittins index by Theorem II.13, part (ii).

If the channel observed at time $t-1$ is in state $i, i < K$, then the PMF of that channel at t is P_i and the PMFs of all other channels are stochastically ordered and are stochastically larger than P_{K-1} and stochastically smaller than P_K by Property II.4. The myopic policy will choose the channel with the stochastically largest PMF (among all channels that are not observed at $t-1$). By Theorem II.13 (ii), the Gittins index of the same channel is the largest among the Gittins indices of all channels that are not observed at $t-1$. By Theorem II.13 (iii), the Gittins index of the channel observed at time $t-1$ is $\nu(P_i) \leq \nu(\pi)$ for all $P_{K-1} \leq_{st} \pi \leq_{st} P_K$. Therefore, the Gittins index chooses the same channel as the myopic policy. From the optimality of the myopic policy, under Conditions (A1)-(A4) (Theorem II.2) and the condition $L = K$, after time $t = 0$ the Gittins index rule is an optimal channel sensing strategy for problem (P1) and (P2). □

Note that, if two channels, say channel 1 and 2 are such that $\pi_0^1, \pi_0^2 \in \{P_1, \ldots, P_{K-1}\}$ then $\pi_0^1, \pi_0^2 \in \Delta(S)P$ and thus, (A2) is satisfied. Nevertheless π_0^1, π_0^2 do not necessarily satisfy the condition $P_{k-1} \leq_{st} \pi_0^i \leq_{st} P_K$ of Theorem II.13. Thus, at $t = 0$, the assertion of Theorem II.13 may not be true for channels 1 and 2, thus the Gittins index rule may not be optimal at time 0.

2.6 MDP Approximation and Numerical Experiments

In this section, we consider the POMDP formulation of Problem (P1) for the finite horizon channel sensing problem, and develop a MDP approximation for the POMDP.

In the POMDP, each channel's PMF $\pi_t^n \in \Delta(S)$. However, not all vectors in $\pi_t^n \in \Delta(S)$ are possible PMFs for the channels. For any time t, if the last time channel n is selected before t is $t-s$, and the state of channel n at $t-s$ is $k \in S$, then the PMF of channel at t is equal to

$$\pi_t^n = P_k P^{s-1}. \tag{2.91}$$

Therefore, the PMF of a channel can always be characterized by a pair (k, s) where $k \in S$, $s \in \mathbb{N} := \{1, 2, 3, \dots\}$. Define $\Delta'(S) := \{P_k P^{s-1} : k \in S, s \in \mathbb{N}\}$. The set $\Delta'(S)$ of possible PMFs is a countable set. We can further approximate it with a finite set $\Delta'_M(S)$ for any number M such that

$$\Delta'_M(S) =: \{P_k P^{s-1} : k \in S, s = 1, \dots, M\}. \tag{2.92}$$

Using the finite approximation $\Delta'_M(S)$ of the set of possible PMFs $\Delta'(S)$, we can construct a finite Markov decision process (MDP) that approximates the POMDP for the channel sensing problem.

MDP_M Approximation:

The state of the system at t is $Z_t = (Z_t^1, Z_t^2, \dots, Z_t^N)$, where $Z_t^n = (Z_t^n(1), Z_t^n(2)) \in S \times \{1, \dots, M\}$ for all n for all t. Based on the action $U_t = 1, 2, \dots, N$ at t, the state

of the system evolves as

$$Z_{t+1}^n(1) = \begin{cases} Z_t^n(1) & \text{if } U_t \neq n, \\ Y_t^n & \text{if } U_t = n, \end{cases} \tag{2.93}$$

$$Z_{t+1}^n(2) = \begin{cases} \min(Z_t^n(2)+1, M) & \text{if } U_t \neq n, \\ 1 & \text{if } U_t = n. \end{cases} \tag{2.94}$$

The reward incurred at time t is given by

$$R(t) = P_{Z_t^{U_t}(1)} P^{(Z_t^{U_t}(2)-1)} R. \tag{2.95}$$

The above finite state MDP_M can be solved by dynamic programming. Define a map $\eta_M : \Delta'(S) \mapsto S \times \{1, \ldots, M\}$ as

$$\eta_M(\pi_t^n) = (k, \min(s, M)) \text{ when } \pi_t^n = P_k P^{s-1}. \tag{2.96}$$

Then, we can construct a separating channel sensing policy $g^M \in \mathcal{G}_s$ from an optimal policy $\tilde{g}^M$ for MDP_M by

$$U_t = g_t^m(\pi_t^1, \pi_t^2, \ldots, \pi_t^N) = \tilde{g}_t^M(\eta_M(\pi_t^1), \eta_M(\pi_t^2), \ldots, \eta_M(\pi_t^N)). \tag{2.97}$$

As M increases, $\lim_{M\to\infty} \Delta'_M(S) = \Delta'(S)$. Therefore, we expect that the performance of g^M will approach the optimal policy as M increases. However, the complexity of solving MDP_M also increases with M because the size of the state space in MDP_M is $(KM)^N$.

In this section, we numerically solve MDP_M for an instance with $N = 3$ channels

and $K = 3$ states over time horizon $T = 10$ where

$$\beta = 1, \tag{2.98}$$

$$P = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}, \tag{2.99}$$

$$R = \begin{bmatrix} 1 & 2 & 3 \end{bmatrix}^T, \tag{2.100}$$

$$\pi_0^1 = \pi_0^2 = \pi_0^3 = P_1 P^{10}. \tag{2.101}$$

Note that, (A2) is not satisfied under the above parameters; therefore, the myopic policy may not be optimal for this instance.

We use simulation to compare the performance of the myopic policy to the performance of g^M for different values of M. In the simulation, we ran the channel sensing problem for 5000 times, and use the average reward over all experiments as the performance criteria of a policy.

Fig. 2.2 shows the performance of the myopic policy g^m, policy g^M, random selection, and the maximum reward if all channels' states are available. It is shown in Fig. 2.2 that the performance of policy g_M increases as M increases. Furthermore, policy g^M outperforms the myopic policy when M is large ($M \geq 3$).

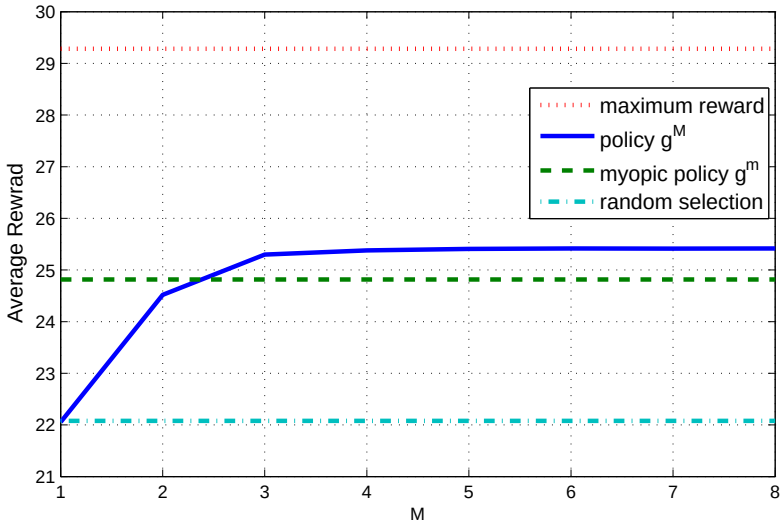

Figure 2.2: The performance of the myopic policy g^m and policy g^M

2.7 Multiple Selection

In this section, we consider the channel sensing problem with multiple selection. That is, the user can select multiple channels to transmit data at each time. We compare the performance of the myopic policy and the policy constructed from a MDP approximation for the multiple selection channel sensing problem.

Suppose the user can select d channels at each time, then the user's transmission decision U_t is a set of channels of size d. Each channel in $U_t \subset \{1, 2, \ldots, N\}$ is used by the user for transmission at t. The reward at t from transmission through the d channels is

$$R_{MS}(t) = \sum_{i \in U_t} R_{X_t^i}. \tag{2.102}$$

We define the multiple selection channel sensing problem $MS(d)$ where the user selects d channels at each time.

Problem $MS(d)$

$$\max_{g \in \mathcal{G}_{MS(d)}} \mathbb{E}^g \left[\sum_{t=0}^{T} \beta^t R_{MS}(t) \right], \tag{2.103}$$

where $\mathcal{G}_{MS(d)}$ is the set of all separated policies from $\Delta(S)^N$ to the set of size d subsets in $\{1, 2, \ldots, N\}$.

When $d = 1$, the above described multiple selection problem $MS(d)$ becomes the single selection channel sensing problem (P1) formulated in Section 2.2.1.

For Problem $MS(d)$, one simple policy is the myopic policy defined below.

Definition II.15. The myopic policy $g^{(m,d)} := (g_0^{(m,d)}, g_1^{(m,d)}, \ldots, g_T^{(m,d)})$ is the policy that selects at each time instant the d channels with the first d largest expected reward; that is, if O_t is the channel ordering such that

$$\pi_t^{O_t(1)} R \geq \pi_t^{O_t(2)} R \geq \cdots \geq \pi_t^{O_t(N)} R \tag{2.104}$$

Then

$$g_t^{(m,d)}(\pi_t) = \{O_t(1), O_t(2), \ldots, O_t(d)\}. \tag{2.105}$$

Using similar approximation ideas from Section 2.6, we can also construct finite MDP approximation for Problem $MS(d)$.

$MS(d) - MDP_M$ *approximation*

The state of the system at t is $Z_t = (Z_t^1, Z_t^2, \ldots, Z_t^N)$, where $Z_t^n = (Z_t^n(1), Z_t^n(2)) \in S \times \{1, \ldots, M\}$ for all n for all t. Based on the action $U_t \subset \{1, 2, \ldots, N\}$ at t, the

state of the system evolves as

$$Z_{t+1}^n(1) = \begin{cases} Z_t^n(1) & \text{if } n \notin U_t \\ Y_t^n & \text{if } n \in U_t \end{cases} \tag{2.106}$$

$$Z_{t+1}^n(2) = \begin{cases} \min(Z_t^n(2)+1, M) & \text{if } n \notin U_t \\ 1 & \text{if } n \in U_t \end{cases} \tag{2.107}$$

The reward incurred at time t is given by

$$R(t) = \sum_{n \in U_t} P_{Z_t^n(1)} P^{(Z_t^n(2)-1)} R. \tag{2.108}$$

Similar to the case of single selection in Section 2.6, from any solution $\tilde{g}^{(M,d)}$ for $MS(d) - MDP_M$, we can construct a separating channel sensing policy $g^{(M,d)} \in \mathcal{G}_s$ by

$$U_t = g_t^{(M,d)}(\pi_t^1, \pi_t^2, \ldots, \pi_t^N) = \tilde{g}_t^{(M,d)}(\eta_M(\pi_t^1), \eta_M(\pi_t^2), \ldots, \eta_M(\pi_t^N)). \tag{2.109}$$

We numerically solved $MS(d) - MDP_M$ for the problem instance in Section 2.6 with double selection ($d = 2$), and use simulation to compare the performance of the myopic policy $g^{m,2}$ to the performance of $g^{(M,2)}$ for different values of M. In the simulation, we ran the multiple selection channel sensing problem for 5000 times, and use the average reward over the 5000 experiments as the performance criteria of a policy.

Fig. 2.3 shows the performance of the myopic policy $g^{(m,2)}$, policy $g^{(M,2)}$, random selection, and the maximum reward if all channels' states are available. Policy $g^{(M,2)}$ performs better than the myopic policy $g^{(m,2)}$ when $M \geq 2$. However, the performance difference between $g^{(M,2)}$ and $g^{(m,2)}$ in the multiple selection problem $MS(d)$ is smaller than that between g^M and g^m in the single selection problem (P1) (see Fig. 2.2).

The simulation results suggest that the performance loss due to the myopic policy, resulting from its comparison with the optimal policy, is smaller when more channels can be selected at each time instant.

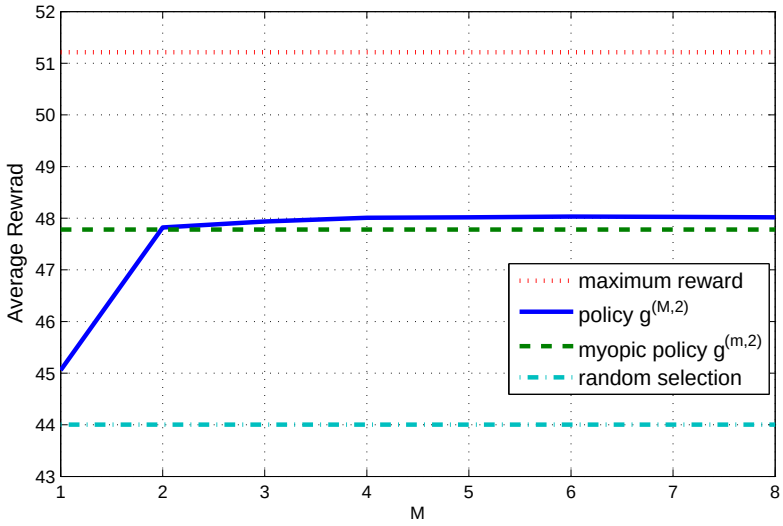

Figure 2.3: The performance of $g^{(m,2)}$ and $g^{(M,2)}$ in Problem $MS(2)$.

2.8 Conclusion

The channel sensing problem we investigated in this chapter arises in communications and many other fields of science and technology, as it is an instance of restless bandit problems. For the single selection problem, we identified conditions sufficient to guarantee the optimality of the myopic policy, and conditions under which the Gittins index rule coincides with the myopic policy (and is optimal). We also developed a MDP approximation that results in a dynamic programming algorithm for computing a near-optimal policy. For the multiple selection problem, we computed a near-optimal policy from a MDP approximation, and compared it with the myopic policy in numerical simulations.

Our results on the optimality of the myopic policy for the single selection problem extend previously known results on the same problem when each channel has two

states. As pointed out in Section 2.3.1, such an extension is non-trivial and requires significant insight into the nature of the problem (so as to identify the appropriate assumptions/conditions), and much more careful and complicated analysis (so as to discover qualitative properties of optimal sensing policies, such as the optimality of the myopic policy).

Our results on the optimality of the Gittins index rule for the single selection problem rely on : (i) the fact that the information state of any channel after $t > 0$ lies stochastically between P_{K-1} and P_K, i.e. $P_{K-1} \leq_{st} \pi \leq_{st} P_K$; and (ii) the fact that $\nu(\hat{\pi}) \geq \nu(\pi)$ whenever $\hat{\pi} \geq_{st} \pi$ and both $\hat{\pi}$ and π are stochastically ordered between P_{K-1} and P_K. We have not been able to prove whether or not the Gittins index rule coincides with the myopic policy when conditions (A1)-(A4) are valid and $L \neq K$ in (A3).

Our simulation results for the multiple selection problem suggest that the performance loss due to the myopic policy, resulting from its comparison with the optimal policy, is smaller when more channels can be selected at each time instant.

CHAPTER III

Decentralized Stochastic Control-Part I: Decentralized Routing

3.1 Introduction

Routing problems to parallel queues arise in many modern technological systems such as communication, transportation and sensor networks. The majority of the literature on optimal routing in parallel queues addresses situations where the information is centralized, either perfect (see [45–57] and references therein) or imperfect (see [58, 59] and references therein). Very few results on optimal routing to parallel queues under decentralized information are currently available. The authors of [60] present a heuristic approach to decentralized routing in parallel queues. In ([61–65] and references therein), decentralized routing policies that stablize queueing networks are considered. The work in [66] presents an optimal policy to a routing problem with a one-unit delay sharing information structure.

In this chapter we investigate a decentralized routing problem in discrete time. We consider a system consisting of two service stations/queues, called Q_1 and Q_2 and two controllers, called C_1 and C_2. Controller C_1 (resp. C_2) is affiliated with service station Q_1 (resp. Q_2). Each station has an infinite size buffer. The processes describing exogenous customer arrivals at each station are independent Bernoulli with

parameter (λ). The random variables describing the service times at each station are independent geometric with parameter (μ). At any time each controller can route one of the customers waiting in its own queue to the other station. Each controller knows perfectly the queue length in its own station, and observes the exogenous arrivals in its own station as well as the arrivals of customers sent from the other station. At the beginning, controller C_1 (resp. C_2) has a probability mass function (PMF) on the number of customers in station Q_2 (resp. Q_1). These PMFs are common knowledge between the controllers. At each time a holding cost is incurred at each station due to the customers waiting at that station. The objective is to determine decentralized routing policies for the two controllers that minimize either the total expected holding cost over a finite horizon or the average cost per unit time over an infinite horizon.

In the above described routing problem, each controller has different information. Furthermore, the control actions/routing decisions of one controller affect the information of the other controller. Thus, the information structure of this decentralized routing problem is non-classical with control sharing (see [67] for non-classical control sharing information structures). Non-classical information structures result in challenging signaling problems (see [3]). Signaling occurs through the routing decisions of the controllers. Signaling is, in essence, a real-time encoding/communication problem within the context of a decision making problem. By sending or not sending a customer from Q_1 (resp. Q_2) to Q_2 (resp. Q_1) controller C_1 (resp. C_2) communicates at each time instant a compressed version of its queue length to C_2 (resp. C_1). For example, by sending a customer from Q_1 to Q_2 at time t, C_1 may signal to C_2 that Q_1's queue length is above a pre-specified threshold l_t. This information allows C_2 to have a better estimate of Q_1's queue length and, therefore, make better routing decisions about the customers in its own queue; the same arguments hold for the signals send (through routing decisions) from C_2 to C_1. Thus, signaling through routing decisions has a triple function: communication, estimation and control.

Within the context of the problems described above, there is enormous number of signaling possibilities. For example, there is an arbitrarily large number of choices of the sequences of pre-specified thresholds $l_1, l_2, \ldots, l_t, \ldots$ and these choices are only a small subset of all the possible sequences of binary partitions of the set of non-negative integers that describe all choices available to C_1 and C_2. All these possibilities result in highly non-trivial decision making problems. It is the presence of signaling that distinguishes the problem formulated in this chapter from all other routing problems in parallel queues investigated so far.

Some basic questions associated with the analysis of this problem are:

What is an information state (sufficient statistic) for each controller? How is signaling incorporated in the evolution/update of the information state? Is there an explicit description of an optimal signaling strategy? We will answer these questions in Section 3.3-3.6 and will discuss them further in Section 3.9.

Organization

The rest of the chapter is organized as follows. In Section 3.2 we present the model for the queueing system and formulate the finite horizon and infinite horizon decentralized routing problems. In Section 3.3 we present structural results for optimal policies. In Section 3.4 we present a specific decentralized routing policy, which we call $\hat{g}$, and state some features associated with its performance. In Section 3.5, we show that when the initial queue lengths in Q_1 and Q_2 are equal, $\hat{g}$ is an optimal policy for the finite horizon decentralized routing problem. In Section 3.6, we show that $\hat{g}$ is an optimal decentralized routing policy for the infinite horizon average cost per unit time problem. In Section 3.7, we consider the system with bursty arrivals and present the decentralized routing policy DR_M. We present numerical examples in Section 3.8 to illustrate the results developed in Section 3.2-3.7. We conclude in Section 3.9.

3.2 System Model and Problem Formulation

System Model

The queueing/service system shown in Figure 3.1, operates in discrete time.

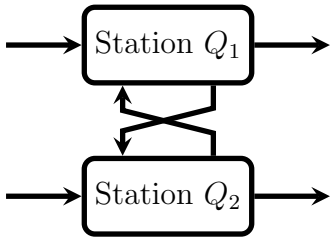

Figure 3.1: The queueing system.

The system consists of two service stations/queues, Q_1 and Q_2 with infinite size buffers. Controllers C_1 and C_2 are affiliated with queues Q_1 and Q_2, respectively. Let X_t^i denote the number of customers waiting, or in service, in $Q_i, i = 1, 2$, at the beginning of time t. Exogenous customer arrivals at $Q_i, i = 1, 2$, occur according to a Bernoulli process $\{A_t^i, t \in \mathbb{Z}_+\}$ with parameter λ. Service times of customers at $Q_i, i = 1, 2$ are described by geometric random variables with parameter μ. We define a Bernoulli process $\{D_t^i, t \in \mathbb{Z}_+\}$ with parameter μ. Then $\{D_t^i 1_{\{X_t^i \neq 0\}}, t \in \mathbb{Z}_+\}$ describes the customer departure process from $Q_i, i = 1, 2$. At any time t, a controller can route one of the customers in its own queue to the other queue. Let U_t^i denote the routing decision of controller C_i at t $(i = 1, 2)$; if $U_t^i = 1$ (resp. 0) one customer (resp. no customer) is routed from Q_i to Q_j $(j \neq i)$. At any time t, each controller $C_i, i = 1, 2$, knows perfectly the number of customers $X_{0:t}^i, i = 1, 2$, in its own queue; furthermore, it observes perfectly the arrival stream $A_{0:t}^i$ to its own queue, and the arrivals due to customers routed to its queue from the other service station up to time $t - 1$, i.e. $U_{0:t-1}^j, j \neq i$. The order of arrivals A_t^i, departures D_t^i and controller decisions U_t^i concerning the routing of customers from one queue to the other is shown in Figure 3.2.

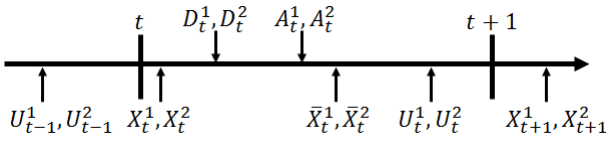

Figure 3.2: The order of variables

The dynamic evolution of the number of customers $X_t^i, i = 1, 2$ is described by

$$X_{t+1}^1 = \overline{X}_t^1 - U_t^1 + U_t^2, \tag{3.1}$$

$$X_{t+1}^2 = \overline{X}_t^2 - U_t^2 + U_t^1, \tag{3.2}$$

where for $i = 1, 2$,

$$\overline{X}_t^i = \left(X_t^i - D_t^i\right)^+ + A_t^i, \tag{3.3}$$

and $(x)^+ := \max(0, x)$. We assume that the initial queue lengths X_0^1, X_0^2 and the processes $\{A_t^1, t \in \mathbb{Z}_+\}$, $\{A_t^2, t \in \mathbb{Z}_+\}$, $\{D_t^1, t \in \mathbb{Z}_+\}$, $\{D_t^2, t \in \mathbb{Z}_+\}$ are mutually independent and their distributions are known by both controllers C_1 and C_2. Let π_0^1 and π_0^2 be the PMFs on the initial queue lengths X_0^1, X_0^2, respectively. At the beginning of time $t = 0$, C_1 (resp. C_2) knows X_0^1 (resp. X_0^2). Furthermore C_1's (resp. C_2's) knowledge of the queue length X_0^2 (resp. X_0^1) at the other station is described by the PMF π_0^2 (resp. π_0^1). The information of controller $C_i, i = 1, 2$, at the moment it makes the decision $U_t^i, t = 0, 1, \ldots,$ is

$$I_t^i := \left\{X_{0:t}^i, A_{0:t}^i, \overline{X}_{0:t}^i, U_{0:t-1}^1, U_{0:t-1}^2, \pi_0^1, \pi_0^2\right\}, i = 1, 2. \tag{3.4}$$

The controllers' routing decisions/control actions U_t^i are generated according to

$$U_t^i = g_t^i\left(I_t^i\right), i = 1, 2, t \in \mathbb{Z}_+, \tag{3.5}$$

where

$$g_t^i : (\mathbb{Z}_+)^{t+1} \times \{0,1\}^{t+1} \times (\mathbb{Z}_+)^{t+1} \times \{0,1\}^t \times \\ \times \{0,1\}^t \times \mathbb{R}^{\mathbb{Z}_+} \times \mathbb{R}^{\mathbb{Z}_+} \mapsto \mathcal{U}_t^i. \tag{3.6}$$

The control action space $\mathcal{U}_t^i$ at time t depends on $\overline{X}_t^i$. Specifically

$$\mathcal{U}_t^i = \begin{cases} \{0\} & \text{when } \overline{X}_t^i = 0, \\ \{0,1\} & \text{otherwise.} \end{cases} \tag{3.7}$$

Define $\mathcal{G}_d$ to be the set of feasible decentralized routing policies; that is

$$\mathcal{G}_d = \{(g^1, g^2) : \quad g^i = (g_0^i, g_1^i, \dots, g_t^i, \dots), i = 1, 2 \\ \text{and } g_t^i \text{ is of form given by (3.5)-(3.6)}\}. \tag{3.8}$$

We study the operation of the system defined in this section, first over a finite horizon, then over an infinite horizon.

3.2.1 The finite horizon problem

For the problem with a finite horizon T, we assume the holding cost incurred by the customers present in Q_i at time $t = 0, 1, \dots, T-1$ is $c_t(X_t^i), i = 1, 2$, where $c_t(\cdot)$ is a convex and increasing function. Then, the objective is to determine decentralized routing policies $g \in \mathcal{G}_d$ so as to minimize

$$J_T^g(\pi_0^1, \pi_0^2) := \mathbb{E}\left[\sum_{t=0}^{T-1} \left(c_t\left(X_t^{1,g}\right) + c_t\left(X_t^{2,g}\right)\right) \middle| \pi_0^1, \pi_0^2\right] \tag{3.9}$$

for any PMFs π_0^1, π_0^2 on the initial queue lengths.

3.2.2 The infinite horizon average cost per unit time problem

For the infinite horizon average cost per unit time problem, we assume the holding cost incurred by the customers present in Q_i at each time is a convex and increasing function $c_t(\cdot) := c(\cdot), i = 1, 2$. Then, the objective is to determine decentralized routing policies $g = (g^1, g^2) \in \mathcal{G}_d$ so as to minimize

$$\begin{aligned} & J^g(\pi_0^1, \pi_0^2) \\ :=& \limsup_{T\to\infty} \frac{1}{T} J_T^g(\pi_0^1, \pi_0^2) \\ =& \limsup_{T\to\infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \left(c\left(X_t^{1,g}\right) + c\left(X_t^{2,g}\right)\right) \middle| \pi_0^1, \pi_0^2 \right] \end{aligned} \tag{3.10}$$

for any PMFs π_0^1, π_0^2 on the initial queue lengths.

3.3 Qualitative Properties of Optimal Policies

In this section we present a qualitative property of an optimal routing policy for both the finite horizon and the infinite horizon problem. For that matter we first introduce the following notation.

We denote by Π_t^1 and Π_t^2 the PMFs on X_t^1 and X_t^2, respectively, conditional on all previous decisions $\{U_{0:t-1}^1, U_{0:t-1}^2\}$. $\Pi_t^i, i = 1, 2$ is defined by

$$\Pi_t^i(x) := \mathbb{P}\left(X_t^i = x | U_{0:t-1}^1, U_{0:t-1}^2\right), x \in \mathbb{Z}_+. \tag{3.11}$$

Similarly, we define the conditional PMFs $\overline{\Pi}_t^1$, $\overline{\Pi}_t^2$ on $\overline{X}_t^1$ and $\overline{X}_t^2$, respectively, as follows.

$$\overline{\Pi}_t^i(x) := \mathbb{P}\left(\overline{X}_t^i = x | U_{0:t-1}^1, U_{0:t-1}^2\right), i = 1, 2, x \in \mathbb{Z}_+. \tag{3.12}$$

Note that for any policy $g \in \mathcal{G}_d$ all the above defined PMFs are functions of $\{U_{0:t-1}^1, U_{0:t-1}^2\}$.

Since both controllers C_1 and C_2 know $\{U^1_{0:t-1}, U^2_{0:t-1}\}$ at time t, the PMFs defined by (3.11)-(3.12) are common knowledge [68] between the controllers.

We take $\overline{X}^i_t, i = 1, 2$, to be station Q_i's state at time t. Combining (3.1)-(3.3) we obtain, for $i = 1, 2$,

$$\begin{aligned}\overline{X}^i_{t+1} =& \left(\overline{X}^i_t - U^i_t + U^j_t - D^i_{t+1}\right)^+ + A^i_{t+1} \\ :=& f^i_t\left(\overline{X}^i_t, U^i_t, U^j_t, W^i_t\right),\end{aligned} \tag{3.13}$$

where the random variables $W^i_t := (A^i_{t+1}, D^i_{t+1}), i = 1, 2, t = 0, 1, \ldots$ are mutually independent.

The holding cost at time $t, t = 0, 1, \ldots$ can be written as

$$\begin{aligned}&\rho_t\left(\overline{X}^1_t, \overline{X}^2_t, U^1_t, U^2_t\right) \\ :=& c_{t+1}\left(\overline{X}^1_t - U^1_t + U^2_t\right) + c_{t+1}\left(\overline{X}^2_t - U^2_t + U^1_t\right) \\ =& c_{t+1}\left(X^1_{t+1}\right) + c_{t+1}\left(X^2_{t+1}\right).\end{aligned} \tag{3.14}$$

Note that for any time horizon T the total expected holding cost due to (3.14) is equivalent to the total expected holding cost defined by (3.9) since for any policy $g \in \mathcal{G}_d$

$$\begin{aligned}&J^g_T(\pi^1_0, \pi^2_0) \\ =& \mathbb{E}\left[\sum_{t=0}^{T-1}\left(c_t\left(X^{1,g}_t\right) + c_t\left(X^{2,g}_t\right)\right)\right] \\ =& \mathbb{E}\left[\sum_{t=0}^{T-2}\left(c_{t+1}\left(X^{1,g}_{t+1}\right) + c_{t+1}\left(X^{2,g}_{t+1}\right)\right)\right] \\ &+ \mathbb{E}\left[c_0\left(X^1_0\right) + c_0\left(X^2_0\right)\right] \\ =& \mathbb{E}\left[\sum_{t=0}^{T-2}\rho_t\left(\overline{X}^{1,g}_t, \overline{X}^{2,g}_t, U^1_t, U^2_t\right)\right] + \mathbb{E}\left[c_0\left(X^1_0\right) + c_0\left(X^2_0\right)\right].\end{aligned} \tag{3.15}$$

With the above notation and definition of system state and instantaneous holding cost, we have a dynamic team problem with non-classical information structure where the common information between the two controllers at any time t is their decisions/control actions up to time $t-1$. This information structure is the control sharing information structure investigated in [67]. Furthermore, the independent assumption we made about the exogenous arrivals and the service processes is the same as the assumptions made about the noise variables in [67]. Therefore, the following Properties III.1-III.3 hold by the results in [67].

Proposition III.1. *For each t, and any given $g_s^1(.), g_s^2(.), s \leq t$, we have*

$$\begin{aligned}&\mathbb{P}\left(I_t^1 = i_t^1, I_t^2 = i_t^2 | U_{0:t-1}^1, U_{0:t-1}^2\right)\\ =&\mathbb{P}\left(I_t^1 = i_t^1 | U_{0:t-1}^1, U_{0:t-1}^2\right)\mathbb{P}\left(I_t^2 = i_t^2 | U_{0:t-1}^1, U_{0:t-1}^2\right).\end{aligned} \tag{3.16}$$

Proof. Same as that of Proposition 2 in [67]. □

Property III.1 says that the two subsystems are independent conditional on past control actions.

Because of Property III.1 and (3.13), each controller $C_i, i = 1, 2$ can generate its decision at any time t by using only its current local state $\overline{X}_t^i$ and past decisions of both controllers. This assertion is established by the following property.

Proposition III.2. *For the routing problems formulated in Section 3.2, without loss of optimality we can restrict attention to routing policies of the form*

$$U_t^1 = g_t^1\left(\overline{X}_t^1, U_{0:t-1}^1, U_{0:t-1}^2\right), \tag{3.17}$$

$$U_t^2 = g_t^2\left(\overline{X}_t^2, U_{0:t-1}^1, U_{0:t-1}^2\right). \tag{3.18}$$

Proof. Same as that of Proposition 1 in [67]. □

Using the common information approach in [13], we can refine the result of Property III.2 as follows.

Proposition III.3. *For the two routing problems formulated in Section 3.2, without loss of optimality we can restrict attention to routing policies of the form*

$$U_t^1 = g_t^1\left(\overline{X}_t^1, \overline{\Pi}_t^1, \overline{\Pi}_t^2\right), \tag{3.19}$$

$$U_t^2 = g_t^2\left(\overline{X}_t^1, \overline{\Pi}_t^1, \overline{\Pi}_t^2\right). \tag{3.20}$$

Proof. Same as that of Theorem 1 in [67]. □

The result of Property III.3 will play a central role in the analysis of the decentralized routing problems formulated in this chapter.

3.4 The Decentralized Policy $\hat{g}$ and Preliminary results

In this section, we specify a decentralized policy $\hat{g}$ and identify an information state for each controller. Furthermore, we develop some preliminary results for both the finite horizon problem and the infinite horizon problem formulated in Section 3.2.

To specify policy $\hat{g}$, we first define the upper bound and lower bound on the support of the PMF, $\Pi_t^i, i = 1, 2$ as

$$UB_t^i := \max(x : \Pi_t^i(x) \neq 0), \tag{3.21}$$

$$LB_t^i := \min(x : \Pi_t^i(x) \neq 0). \tag{3.22}$$

$$UB_t := \max(UB_t^1, UB_t^2), \tag{3.23}$$

$$LB_t := \min(LB_t^1, LB_t^2). \tag{3.24}$$

Similarly, we define the bounds on the support of the PMF, $\overline{\Pi}_t^i, i = 1, 2$ as

$$\overline{UB}_t^i := \max(x : \overline{\Pi}_t^i(x) \neq 0), \tag{3.25}$$

$$\overline{LB}_t^i := \min(x : \overline{\Pi}_t^i(x) \neq 0), \tag{3.26}$$

$$\overline{UB}_t := \max(\overline{UB}_t^1, \overline{UB}_t^2), \tag{3.27}$$

$$\overline{LB}_t := \min(\overline{UB}_t^1, \overline{UB}_t^2). \tag{3.28}$$

Using the above bounds, we specify the policy $\hat{g} := (\hat{g}^1, \hat{g}^2)$ as follows:

$$U_t^i = \hat{g}_t^i \left(\overline{X}_t^i, \overline{UB}_t, \overline{LB}_t \right) = \begin{cases} 1, \text{ when } \overline{X}_t^i \geq TH_t, \\ 0, \text{ when } \overline{X}_t^i < TH_t, \end{cases} \tag{3.29}$$

where

$$TH_t = \frac{1}{2} \left(\overline{UB}_t + \overline{LB}_t \right). \tag{3.30}$$

Under $\hat{g}$, each controller routes a customer to the other queue when $\overline{X}_t^i, i = 1, 2$, the queue length of its own station at the time of decision, is greater than or equal to the threshold given by (3.30).

Note that this decentralized routing policy $\hat{g}$ is indeed of the form asserted by Property III.3 since the upper and lower bounds $\overline{UB}_t$ and $\overline{LB}_t$ are both functions of the PMFs $\overline{\Pi}_t^1, \overline{\Pi}_t^2$. Therefore, the threshold TH_t, as a function of $\overline{\Pi}_t^1, \overline{\Pi}_t^2$, is common knowledge between the controllers. Using the common information, each controller can compute the threshold according to (3.30) individually, and $\hat{g}$ can be implemented in a decentralized manner.

Under policy $\hat{g}$, the evolution of the bounds defined by (3.23)-(3.28) are determined by the following lemma.

Lemma III.4. *At any time t we have*

$$\overline{UB}_t^{\hat{g}} = UB_t^{\hat{g}} + 1, \quad \overline{LB}_t^{\hat{g}} = \left(LB_t^{\hat{g}} - 1\right)^+. \tag{3.31}$$

When $(U_t^{1,\hat{g}}, U_t^{2,\hat{g}}) = (0,0)$

$$UB_{t+1}^{\hat{g}} = \lceil TH_t \rceil - 1, \quad LB_{t+1}^{\hat{g}} = \overline{LB}_t^{\hat{g}} \tag{3.32}$$

When $(U_t^{1,\hat{g}}, U_t^{2,\hat{g}}) = (1,1)$

$$UB_{t+1}^{\hat{g}} = \overline{UB}_t^{\hat{g}}, \quad LB_{t+1}^{\hat{g}} = \lceil TH_t \rceil \tag{3.33}$$

When $(U_t^{i,\hat{g}}, U_t^{j,\hat{g}}) = (1,0), i = 1,2, j \neq i$

$$UB_{t+1}^{\hat{g}} = \max\left(\overline{UB}_t^{i,\hat{g}} - 1, \lceil TH_t \rceil\right) \tag{3.34}$$

$$LB_{t+1}^{\hat{g}} = \min\left(\overline{LB}_t^{j,\hat{g}} + 1, \lceil TH_t \rceil - 1\right) \tag{3.35}$$

where $\lfloor x \rfloor$ = *maximum integer* $\leq x$, *and* $\lceil x \rceil$ = *minimum integer* $\geq x$.

Proof. See Appendix B. □

Corollary III.5 below follows directly form (3.31)-(3.35) in Lemma III.4.

Corollary III.5. *Under policy* $\hat{g}$,

$$\begin{aligned} & UB_{t+1}^{\hat{g}} - LB_{t+1}^{\hat{g}} \\ \leq & \begin{cases} \left\lceil \frac{1}{2}\left(UB_t^{\hat{g}} - LB_t^{\hat{g}}\right) \right\rceil & \text{when } (U_t^{1,\hat{g}}, U_t^{2,\hat{g}}) = (0,0), \\ UB_t^{\hat{g}} - LB_t^{\hat{g}} & \text{otherwise.} \end{cases} \end{aligned} \tag{3.36}$$

Moreover, if $UB_{t_0}^{\hat{g}} - LB_{t_0}^{\hat{g}} \leq 1$ *for some time* t_0*, then*

$$\left(UB_t^{\hat{g}} - LB_t^{\hat{g}}\right) \leq 1 \text{ for all } t \geq t_0. \tag{3.37}$$

Corollary III.5 shows that the difference between the highest possible number of customers in Q_1 or Q_2 and the lowest possible number of customers in Q_1 or Q_2 is non-increasing under the policy $\hat{g}$. Furthermore, the difference is reduced by half when there is no customer routed from one queue to another one.

3.5 The finite horizon problem

In this section, we consider the finite horizon problem formulated in Section 3.2.1, under the additional condition $X_0^1 = X_0^2 = x_0$, where x_0 is arbitrary but fixed, and is common knowledge between C_1 and C_2.

3.5.1 Analysis

The main result of this section asserts that the policy $\hat{g}$ defined in Section 3.4 is optimal.

Theorem III.6. *When* $X_0^1 = X_0^2 = x_0$ *and* x_0 *is common knowledge between* C_1 *and* C_2*, the policy* $\hat{g}$ *given by (3.29)-(3.30) is optimal for the finite horizon decentralized routing problem formulated in Section 3.2.1, that is*

$$J_T^{\hat{g}}(x_0, x_0) \leq J_T^{g}(x_0, x_0) \tag{3.38}$$

for any feasible policy $g \in \mathcal{G}_d$ *and any initial queue length* x_0*.*

Before proving Theorem III.6, we note that when $X_0^1 = X_0^2 = x_0$ Corollary III.5

implies that

$$UB_t^{\hat{g}} - LB_t^{\hat{g}} \leq 1 \text{ for all } t \geq 0. \tag{3.39}$$

Equation (3.39) says that the difference between the highest possible number of customers in Q_1 or Q_2 and the lowest possible number of customers in Q_1 or Q_2 is less than or equal to 1 under policy $\hat{g}$. This property means that $\hat{g}$ controls the length of the joint support of the PMFs $\overline{\Pi}_t^1, \overline{\Pi}_t^2$ and balances the lengths of the two queues. A direct consequence of (3.39) is the following corollary.

Corollary III.7. *At any time t, we have*

$$\left\lfloor \frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}}) \right\rfloor = \min\left(X_t^{1,\hat{g}}, X_t^{2,\hat{g}}\right), \tag{3.40}$$

$$\left\lceil \frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}}) \right\rceil = \max\left(X_t^{1,\hat{g}}, X_t^{2,\hat{g}}\right). \tag{3.41}$$

As pointed out above, the policy $\hat{g}$ balances the lengths of the two queues. This balancing property suggests that the throughput of the system due to $\hat{g}$ is high and the total number of customers in the system is low. This is established by the following lemma.

Lemma III.8. *Under the assumption $X_0^1 = X_0^2 = x_0$, where x_0 is common knowledge, for any policy g of the form described by (3.19)-(3.20), we have*

$$X_t^{1,\hat{g}} + X_t^{2,\hat{g}} \leq_{st} X_t^{1,g} + X_t^{2,g}, \tag{3.42}$$

where $Z_1 \leq_{st} Z_2$ means that the r.v. Z_1 is stochastically smaller than the r.v. Z_2, that is, for any $a \in \mathbb{R}$, $\mathbb{P}(Z_1 \geq a) \leq \mathbb{P}(Z_2 \geq a)$ (see [31]).

Proof. See Appendix B. □

Using Lemma III.8, we now prove Theorem III.6.

Proof of Theorem III.6. For any feasible policy g, since the functions $c_t, t = 0, 1, ..., T$, are convex, we have at any time t

$$\begin{aligned}
&\mathbb{E}\left[c_t\left(X_t^{1,g}\right) + c_t\left(X_t^{2,g}\right)\right] \\
\geq &\mathbb{E}\left[c_t\left(\left\lfloor\frac{1}{2}(X_t^{1,g} + X_t^{2,g})\right\rfloor\right) + c_t\left(\left\lceil\frac{1}{2}(X_t^{1,g} + X_t^{2,g})\right\rceil\right)\right]. \qquad (3.43)
\end{aligned}$$

Furthermore, using Lemma III.8 and the fact that $c_t(\cdot)$ is increasing, we get

$$\begin{aligned}
&\mathbb{E}\left[c_t\left(\left\lfloor\frac{1}{2}(X_t^{1,g} + X_t^{2,g})\right\rfloor\right) + c_t\left(\left\lceil\frac{1}{2}(X_t^{1,g} + X_t^{2,g})\right\rceil\right)\right] \\
\geq &\mathbb{E}\left[c_t\left(\left\lfloor\frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}})\right\rfloor\right) + c_t\left(\left\lceil\frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}})\right\rceil\right)\right] \\
= &\mathbb{E}\left[c_t\left(\min(X_t^{1,\hat{g}}, X_t^{2,\hat{g}})\right) + c_t\left(\max(X_t^{1,\hat{g}}, X_t^{2,\hat{g}})\right)\right] \\
= &\mathbb{E}\left[c_t\left(X_t^{1,\hat{g}}\right) + c_t\left(X_t^{2,\hat{g}}\right)\right]. \qquad (3.44)
\end{aligned}$$

The inequality in (3.44) is true because $X_t^{1,g} + X_t^{2,g} \leq_{st} X_t^{1,\hat{g}} + X_t^{2,\hat{g}}$ (Lemma III.8) and $c_t(\cdot)$ is increasing. The first equality in (3.44) follows from Corollary III.7. Combining (3.43) and (3.44) we obtain, for any t,

$$\mathbb{E}\left[c_t\left(X_t^{1,g}\right) + c_t\left(X_t^{2,g}\right)\right] \geq \mathbb{E}\left[c_t\left(X_t^{1,\hat{g}}\right) + c_t\left(X_t^{2,\hat{g}}\right)\right]. \qquad (3.45)$$

The optimality of policy $\hat{g}$ follows from (3.9) and (3.45). □

3.5.2 Comparison to the performance under centralized information

We compare now the performance of the optimal decentralized policy $\hat{g}$ to the performance of the queueing system under centralized information. The results of this comparison will be useful when we study the infinite horizon problem in Section 3.6.

Consider a centralized controller who has all the information I_t^1 and I_t^2 at each

time t. Then, the set $\mathcal{G}_c$ of feasible routing policies of the centralized controller is

$$\mathcal{G}_c := \{(g^1, g^2) : \quad g^i = (g_0^i, g_1^i, \dots, g_t^i, \dots), i = 1, 2 \\ \text{and } U_t^i = g_t^i(I_t^1, I_t^2)\}. \tag{3.46}$$

By the definition, $\mathcal{G}_d \subset \mathcal{G}_c$. This means that the centralized controller can simulate any decentralized policy $g \in \mathcal{G}_d$ adopted by controllers C_1 and C_2. Therefore, for any initial PMFs π_0^1, π_0^2

$$\inf_{g \in \mathcal{G}_c} J_T^g(\pi_0^1, \pi_0^2) \leq \inf_{g \in \mathcal{G}_d} J_T^g(\pi_0^1, \pi_0^2) \tag{3.47}$$

$$\inf_{g \in \mathcal{G}_c} J^g(\pi_0^1, \pi_0^2) \leq \inf_{g \in \mathcal{G}_d} J^g(\pi_0^1, \pi_0^2). \tag{3.48}$$

When $X_0^1 = X_0^2 = x_0$, Lemma III.8 and Theorem III.6 show that the cost given by $\hat{g}$ is smaller than the cost given by any policy $g \in \mathcal{G}_d$. Furthermore we have:

Lemma III.9. *Under the assumption $X_0^1 = X_0^2 = x_0$, where x_0 is common knowledge, we have*

$$X_t^{1,\hat{g}} + X_t^{2,\hat{g}} \leq_{st} X_t^{1,g} + X_t^{2,g}, \tag{3.49}$$

for any $g \in \mathcal{G}_c$, and

$$J_T^{\hat{g}}(x_0, x_0) \leq \inf_{g \in \mathcal{G}_c} J_T^g(x_0, x_0). \tag{3.50}$$

for any $g \in \mathcal{G}_c$.

Proof. The proof of (3.49) is the same as the proof of Lemma III.8, and the proof of (3.50) is the same as the proof of Theorem III.6. □

Since $\hat{g}$ is a decentralized policy, (3.47) and Lemma III.9 imply that

$$J_T^{\hat{g}}(x_0, x_0) = \inf_{g \in \mathcal{G}_d} J_T^g(x_0, x_0) = \inf_{g \in \mathcal{G}_c} J_T^g(x_0, x_0). \tag{3.51}$$

Equation (3.51) shows that when $X_0^1 = X_0^2 = x_0$ and x_0 is common knowledge between $C_1$and C_2, policy $\hat{g}$ achieves the same performance as any centralized optimal policy.

3.5.3 The Case of Different Initial Queue Lengths

When $X_0^1 \neq X_0^2$, the policy $\hat{g}$ is not necessarily optimal for the finite horizon problem.

Consider an example where the horizon $T = 1$ (two-step horizon), $\lambda = 0.1, \mu = 0.5$ and

$$\mathbb{P}\left(X_0^1 = 3\right) = 1, \tag{3.52}$$

$$\mathbb{P}\left(X_0^2 = 1\right) = 0.9, \quad \mathbb{P}\left(X_0^2 = 5\right) = 0.1, \tag{3.53}$$

that is,

$$\pi_0^1 = (0, 0, 0, 1, 0, 0, 0, \dots), \tag{3.54}$$

$$\pi_0^2 = (0, 0.9, 0, 0, 0, 0.1, 0, \dots), \tag{3.55}$$

where π_0^1, π_0^2 denote the initial PMFs on the lengths of the queues.
Then, $\overline{\Pi}_0^1, \overline{\Pi}_0^2$ and the threshold TH_0 are

$$\overline{\Pi}_0^1 = (0, 0, 0.5, 0.4, 0.1, 0, 0, \dots), \tag{3.56}$$

$$\overline{\Pi}_0^2 = (0.45, 0.36, 0.09, 0, 0.05, 0.04, 0.01, \dots), \tag{3.57}$$

$$TH_0 = \frac{1}{2}(6 + 0) = 3. \tag{3.58}$$

Consider the cost functions $c_0(x) = 0$ and $c_1(x) = x^2$. Then, we have

$$\begin{aligned} & J^g(\pi_0^1, \pi_0^2) \\ =& \mathbb{E}\left[\left(X_1^{1,g}\right)^2 + \left(X_1^{2,g}\right)^2\right] \\ =& \mathbb{E}\left[\left(\overline{X}_0^1 - U_0^{1,g} + U_0^{2,g}\right)^2 + \left(\overline{X}_0^2 - U_0^{2,g} + U_0^{1,g}\right)^2\right]. \end{aligned} \tag{3.59}$$

Using (3.56)-(3.58) and the specification of the policy $\hat{g}$, we can compute the expected cost due to $\hat{g}$. It is

$$J^{\hat{g}}(\pi_0^1, \pi_0^2) = 8.48. \tag{3.60}$$

Consider now another policy $\tilde{g}$ described below. For $i = 1, 2, i \neq j$,

$$U_t^{i,\tilde{g}} = \tilde{g}_t\left(\overline{X}_t^i, \overline{\Pi}_t^1, \overline{\Pi}_t^2\right) = \begin{cases} 1, \text{ when } \overline{X}_t^i \geq \mathbb{E}\left[\overline{X}_t^j | \overline{\Pi}_t^j\right], \\ 0, \text{ when } \overline{X}_t^i < \mathbb{E}\left[\overline{X}_t^j | \overline{\Pi}_t^j\right], \end{cases} \tag{3.61}$$

Then, from (3.56)-(3.57) and (3.61) we get

$$U_0^{1,\tilde{g}} = \begin{cases} 1, \text{ when } \overline{X}_0^1 \geq 1, \\ 0, \text{ when } \overline{X}_0^1 < 1, \end{cases} \tag{3.62}$$

$$U_0^{2,\tilde{g}} = \begin{cases} 1, \text{ when } \overline{X}_0^2 \geq 2.6, \\ 0, \text{ when } \overline{X}_0^2 < 2.6, \end{cases} \tag{3.63}$$

Therefore, the expected cost due to the policy $\tilde{g}$ is given by

$$J^{\tilde{g}}(\pi_0^1, \pi_0^2) = 8.28. \tag{3.64}$$

Since $J^{\tilde{g}}(\pi_0^1, \pi_0^2) = 8.28 < 8.48 = J^{\hat{g}}(\pi_0^1, \pi_0^2)$, policy $\hat{g}$ is not optimal.

In this example, each controller has only one decision to make, the decision at

time 0. As a result, signaling does not provide any advantages to the controllers, and that is why the policy $\hat{g}$ is not the best policy.

3.6 Infinite horizon

We consider the infinite horizon decentralized routing problem formulated in Section 3.2.2, and make the following additional assumptions.

Assumption III.10. $\mu > \lambda$.

Assumption III.11. *The initial PMFs π_0^1, π_0^2 are finitely supported and common knowledge between controllers C_1 and C_2. i.e. there exists $M < \infty$ such that $\pi_0^1(x) = \pi_0^2(x) = 0$ for all $x > M$.*

Let g_0 denote the open-loop policy that does not do any routing, that is, at any time t

$$U_t^{1,g_0} = U_t^{2,g_0} = 0. \tag{3.65}$$

Assumption III.12.

$$\lim_{T\to\infty} \frac{1}{T} J_T^{g_0}(\pi_0^1, \pi_0^2) := J^{g_0} < \infty \quad a.s., \tag{3.66}$$

where J^{g_0} is a constant that denotes the infinite horizon average cost per unit time due to policy g_0.

Remark III.13. Due to policy g_0, the queue length $\{X_t^{i,g_0}, t \in \mathbb{Z}_+\}, i = 1, 2$ is a positive recurrent birth and death chain with arrival rate λ and departure rate $\mu 1_{\{X_t^{g_0,i} \neq 0\}}$. Therefore, as $T \to \infty$, the average cost per unit time converges to a constant a.s. if the expected cost under the stationary distribution of the process is finite (see [69, chap. 3]). Assumption III.12 is equivalent to the assumption that the expected cost is finite under the stationary distribution of the controlled queue lengths.

We proceed to analyze the infinite horizon average cost per unit time for the model of Section 3.2 under Assumptions III.10-III.12.

3.6.1 Analysis

When $X_0^1 \neq X_0^2$, the policy $\hat{g}$, defined in Section 3.4, is not necessarily optimal for the finite horizon problem (see the example in Section 3.5.3). Nevertheless, the policy $\hat{g}$ still attempts to balance the queues. Given enough time, policy $\hat{g}$ may be able to balance the queue lengths even if they are not initially balanced. In this section we show that this is indeed the case.

Specifically, we prove the optimality of policy $\hat{g}$ for the infinite horizon average cost per unit time problem, as stated in the following theorem which is the main result of this section.

Theorem III.14. *Under Assumptions III.10-III.12, the policy $\hat{g}$, described by (3.29)-(3.30), is optimal for the infinite horizon average cost per unit time problem formulated in Section 3.2.2.*

To establish the assertion of Theorem III.14 we proceed in four steps. In the first step we show that the infinite horizon average cost per unit time due to policy $\hat{g}$ is bounded above by the cost of the uncontrolled queues (i.e. the cost due to policy g_0). In the second step we show that under policy $\hat{g}$ the queues are eventually balanced, i.e. the queue lengths can differ by at most one. In the third step we derive a result that connects the performance of policy $\hat{g}$ under the initial PMFs $(0, 0)$ to the performance of the optimal policy under any arbitrary initial PMFs π_0^1, π_0^2 on queues Q_1 and Q_2. In the forth step we establish the optimality of policy $\hat{g}$ based on the results of steps one, two and three.

Step 1

We prove that $J^{\hat{g}}(\pi_0^1, \pi_0^2) \leq J^{g_0}$. To do this, we first establish some preliminary results that appear in Lemmas III.15 and III.16.

Lemma III.15. *There exists processes* $\{Y_t^1, t \in \mathbb{Z}_+\}$ *and* $\{Y_t^2, t \in \mathbb{Z}_+\}$ *such that*

$$\{Y_t^i, t \in \mathbb{Z}_+\} \text{ has the same distribution as } \{X_t^{i,g_0}, t \in \mathbb{Z}_+\} \tag{3.67}$$

for $i = 1, 2$*, and for all times* t

$$X_t^{1,\hat{g}} + X_t^{2,\hat{g}} \leq Y_t^1 + Y_t^2 \quad a.s., \tag{3.68}$$

$$\max_i \left(X_t^{i,\hat{g}}\right) \leq \max_i \left(Y_t^i\right) \quad a.s. \tag{3.69}$$

Proof. See Appendix B. □

Lemma III.15 means that the uncontrolled queue lengths are longer than the queue lengths under policy $\hat{g}$ in a stochastic sense. Note that (3.68) and (3.69) are not true if $Y_t^i, i = 1, 2$, is replaced by $X_t^{i,g_0}, i = 1, 2$, as the following example shows.

Example

When $X_t^{1,g_0} = 4, X_t^{2,g_0} = 6$ and $X_t^{1,\hat{g}} = X_t^{2,\hat{g}} = 5$, the analogues of (3.68) and (3.69) where Y_t^i are replaced by $X_t^{i,g_0}, i = 1, 2$ are

$$X_t^{1,\hat{g}} + X_t^{2,\hat{g}} = X_t^{1,g_0} + X_t^{2,g_0} = 10, \tag{3.70}$$

$$\max_i \left(X_t^{i,\hat{g}}\right) = 5 \leq 6 = \max_i \left(X_t^{i,g_0}\right). \tag{3.71}$$

However, if $A_{t+1}^1 = 1, A_{t+1}^2 = 0$ and $D_{t+1}^1 = 0, D_{t+1}^2 = 1$ we get $X_{t+1}^{1,g_0} = X_{t+1}^{2,g_0} = 5$ and $X_{t+1}^{1,\hat{g}} = 6, X_{t+1}^{2,\hat{g}} = 4$, then

$$\max_i \left(X_{t+1}^{i,\hat{g}}\right) = 6 > 5 = \max_i \left(X_{t+1}^{i,g_0}\right), \tag{3.72}$$

and the analogue of (3.69), when Y_t^i is replaced by $X_t^{i,g_0}, i = 1, 2$, does not hold.

The stochastic dominance relation asserted by Lemma III.15 implies that the instantaneous cost under policy $\hat{g}$ is almost surely no greater than the instantaneous cost due to policy g_0. This implication is made precise by the following lemma.

Lemma III.16. *The processes* $\{Y_t^1, t \in \mathbb{Z}_+\}$ *and* $\{Y_t^2, t \in \mathbb{Z}_+\}$ *defined in Lemma III.15 are such that at any time* t

$$c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right) \leq c\left(Y_t^1\right) + c\left(Y_t^2\right) \quad a.s. \tag{3.73}$$

Proof. See Appendix B. □

In order to apply the result of Step 1 as the time horizon goes to infinity, we need the following result on the convergence of the cost due to $\{Y_t^1, t \in \mathbb{Z}_+\}$ and $\{Y_t^2, t \in \mathbb{Z}_+\}$.

Lemma III.17. *Let* $\{Y_t^1, t \in \mathbb{Z}_+\}$ *and* $\{Y_t^2, t \in \mathbb{Z}_+\}$ *be the processes defined in Lemma III.15. Let* W_T *denote*

$$W_T := \frac{1}{T} \sum_{t=0}^{T-1} \left(c(Y_t^1) + c(Y_t^2)\right). \tag{3.74}$$

Under Assumptions III.11 and III.12,

$$\lim_{T\to\infty} W_T = J^{g_0} \quad a.s. \tag{3.75}$$

Moreover, $\{W_T, T = 1, 2, \dots\}$ *is uniformly integrable, so it also converges in expectation.*

Proof. See Appendix B. □

A direct consequence of Lemmas III.15, III.16 and III.17 is the following.

Corollary III.18. *If* $\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(c\left(X^{1,\hat{g}}\right) + c\left(X^{2,\hat{g}}\right)\right)$ *converges a.s., then,*

$$\frac{1}{T} \sum_{t=0}^{T-1} \left(c\left(X^{1,\hat{g}}\right) + c\left(X^{2,\hat{g}}\right)\right) \longrightarrow J^{\hat{g}}(\pi_0^1, \pi_0^2) \tag{3.76}$$

in expectation and a.s. as $T \to \infty$*. Furthermore,*

$$J^{\hat{g}}(\pi_0^1, \pi_0^2) \leq J^{g_0} < \infty. \tag{3.77}$$

Proof. See Appendix B. □

Step 2

We prove that under policy $\hat{g}$ the queues are eventually balanced. For this matter we first establish some preliminary results that appear in Lemmas III.19 and III.20.

Lemma III.19. *Let* T_0 *be a stopping time with respect to the process* $\{X_t^{1,\hat{g}}, X_t^{2,\hat{g}}, t \in \mathbb{Z}_+\}$*. Define the process* $\{S_t = S_t^{\hat{g}}, t \geq T_0 + 1\}$ *as follows.*

$$S_{T_0+1} := X_{T_0+1}^{1,\hat{g}} + X_{T_0+1}^{2,\hat{g}} \tag{3.78}$$

$$\begin{aligned} S_{t+1} :=& S_t - D_t^1 - D_t^2 + A_t^1 + A_t^2 \\ &+ 1_{\{S_t=1\}} \left(1_{\left\{X_t^{1,\hat{g}}=0\right\}}(D_t^1 - D_t^2) + D_t^2\right) \\ &+ 1_{\{S_t=0\}} \left(D_t^1 + D_t^2\right) \end{aligned} \tag{3.79}$$

If $\mu > \lambda > 0$*, then* $\{S_t, t \geq T_0 + 1\}$ *is an irreducible positive recurrent Markov chain.*

Proof. See Appendix B. □

Lemma III.19 holds for arbitrary stopping time T_0 with respect to $\{X_t^{1,\hat{g}}, X_t^{2,\hat{g}}, t \in \mathbb{Z}_+\}$. By appropriately selecting T_0 we will show later that S_t is coupled with $X_t^{1,\hat{g}} + X_t^{2,\hat{g}}$, i.e. for all $t > T_0$, $S_t = X_t^{1,\hat{g}} + X_t^{2,\hat{g}}$. This result along with the fact that the

process $\{S_t, t \geq T_0 + 1\}$ is an irreducible positive recurrent Markov chain will allow us to analyze the cost due to policy $\hat{g}$.

Lemma III.20. *Under policy $\hat{g}$,*

$$\mathbb{P}\left(\left(U_t^{1,\hat{g}}, U_t^{2,\hat{g}}\right) = (0,0) \quad i.o.\right) = 1. \tag{3.80}$$

Proof. See Appendix B. □

Lemma III.20 means that the event { there exists $t_0 < \infty$ such that at least one of the queue lengths is above the threshold defined by (3.30) for all $t > t_0$ } can not happen. The idea of Lemma III.20 is the following. If one of the queues, say Q_1, has length above the threshold, hence above the lower bound $LB_t^{\hat{g}}$, then, the length of Q_2 does not decrease, because under policy $\hat{g}$, Q_2 receives one customer from Q_1 and has at most one departure at this time. Therefore, both queue lengths at the next time are bounded below by the current lower bound $LB_t^{\hat{g}}$. When at least one of the queue lengths is above the threshold for all $t > t_0$, the queue lengths are bounded below by $LB_{t_0}^{\hat{g}}$ for all $t > t_0$. This kind of lower bound can not exist if the total arrival rate 2λ to the system is less than the total departure rate 2μ from the system.

Lemma III.20 and Corollary III.5 in Section 3.4 can be used to establish that under policy $\hat{g}$ the queues are eventually balanced. This is shown in the corollary below.

Corollary III.21. *Let*

$$T_0 := \inf\{t : UB_t^{\hat{g}} - LB_t^{\hat{g}} \leq 1\}. \tag{3.81}$$

Then

$$\mathbb{P}(T_0 < \infty) = 1, \tag{3.82}$$

$$\left(UB_t^{\hat{g}} - LB_t^{\hat{g}}\right) \leq 1 \text{ for all } t \geq T_0. \tag{3.83}$$

Step 3

We compare the finite horizon cost $J_T^{\hat{g}}(0,0)$ (respectively, the infinite horizon cost $J^{\hat{g}}(0,0)$) due to policy $\hat{g}$ under initial PMFs $(0,0)$ to the minimum finite horizon cost $\inf_{g\in\mathcal{G}_d} J_T^g(\pi_0^1,\pi_0^2)$ (respectively, the minimum infinite horizon cost $\inf_{g\in\mathcal{G}_d} J^g(\pi_0^1,\pi_0^2)$) under arbitrary initial PMFs (π_0^1,π_0^2).

Lemma III.22. *For any finite time T and any initial PMFs π_0^1,π_0^2.*

$$J_T^{\hat{g}}(0,0) = \inf_{g\in\mathcal{G}_c} J_T^g(0,0) \leq \inf_{g\in\mathcal{G}_c} J_T^g(\pi_0^1,\pi_0^2) \leq \inf_{g\in\mathcal{G}_d} J_T^g(\pi_0^1,\pi_0^2), \tag{3.84}$$

and

$$J^{\hat{g}}(0,0) = \inf_{g\in\mathcal{G}_c} J^g(0,0) \leq \inf_{g\in\mathcal{G}_c} J^g(\pi_0^1,\pi_0^2) \leq \inf_{g\in\mathcal{G}_d} J^g(\pi_0^1,\pi_0^2). \tag{3.85}$$

Proof. See Appendix B. □

Lemma III.22 states that the minimum cost achieved when the queues are initially empty is smaller than the minimum cost obtained when the system's initial condition is given by arbitrary PMFs on the lengths of queues Q_1 and Q_2. This result is established through the use of the corresponding centralized information system that is discussed in Section 3.5.2.

Step 4

Based on the results of Steps 1, 2 and 3 we now establish the optimality of policy $\hat{g}$ for the infinite horizon average cost per unit time problem formulated in Section 3.2.2. First, we outline the key ideas in the proof of Theorem III.14, then we present its proof. Step 2 ensures that policy $\hat{g}$ eventually (in finite time) balances the queues. Step 1 ensures that the cost $J^{\hat{g}}(\pi_0^1, \pi_0^2)$ is finite. These two results together imply that the cost due to policy $\hat{g}$ is the same as the cost incurred after the queues are balanced. Furthermore, we show that the cost of policy $\hat{g}$ is independent of the initial PMFs on the queue lengths. Then, the result of Step 3 together with the results on the finite horizon problem establish the optimality of policy $\hat{g}$.

Proof of Theorem III.14. Define T_0 to be the first time when the length of the joint support of PMFs $\Pi_t^{1,\hat{g}}, \Pi_t^{2,\hat{g}}$ is no more than 1. That is

$$T_0 = \inf\{t : UB_t^{\hat{g}} - LB_t^{\hat{g}} \leq 1\}. \tag{3.86}$$

The random variable T_0 is a stopping time with respect to the process $\{X_t^{1,\hat{g}}, X_t^{2,\hat{g}}, t \in \mathbb{Z}_+\}$. From Corollary III.21 we have

$$\mathbb{P}(T_0 < \infty) = 1, \tag{3.87}$$

$$UB_t^{\hat{g}} - LB_t^{\hat{g}} \leq 1 \text{ for all } t \geq T_0. \tag{3.88}$$

Furthermore, for all $t \geq T_0$

$$\left| X_t^{1,\hat{g}} - X_t^{2,\hat{g}} \right| \leq UB_t^{\hat{g}} - LB_t^{\hat{g}} \leq 1. \tag{3.89}$$

Consider the process $\{S_t, t \geq T_0 + 1\}$ defined by (3.78) and (3.79) (in Lemma III.19).

We claim that for all $t \geq T_0 + 1$

$$X_t^{1,\hat{g}} + X_t^{2,\hat{g}} = S_t. \tag{3.90}$$

We prove the claim in Appendix B. Suppose the claim is true. Since $\left|X_t^{1,\hat{g}} - X_t^{2,\hat{g}}\right| \leq 1$ for all $t \geq T_0 + 1$, the instantaneous cost at time $t \geq T_0 + 1$ is equal to

$$\begin{aligned}
& c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right) \\
=& c\left(\left\lfloor \frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}}) \right\rfloor\right) + c\left(\left\lceil \frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}}) \right\rceil\right) \\
=& c\left(\left\lfloor \frac{1}{2}S_t^{\hat{g}} \right\rfloor\right) + c\left(\left\lceil \frac{1}{2}S_t^{\hat{g}} \right\rceil\right).
\end{aligned} \tag{3.91}$$

Then, the average cost per unit time due to policy $\hat{g}$ is given by

$$\begin{aligned}
& \frac{1}{T}\sum_{t=0}^{T-1}\left(c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right)\right) \\
=& \frac{1}{T}\sum_{t=0}^{T_0}\left(c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right)\right) \\
& + \frac{1}{T}\sum_{t=T_0+1}^{T-1}\left(c\left(\left\lfloor \frac{1}{2}S_t^{\hat{g}} \right\rfloor\right) + c\left(\left\lceil \frac{1}{2}S_t^{\hat{g}} \right\rceil\right)\right).
\end{aligned} \tag{3.92}$$

Since $T_0 < \infty\, a.s.$, we obtain

$$\begin{aligned}
&\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right)\right) \\
&= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T_0} \left(c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right)\right) \\
&\quad + \lim_{T\to\infty} \frac{1}{T} \sum_{t=T_0(+1}^{T-1} \left(c\left(\left\lfloor \frac{1}{2} S_t^{\hat{g}} \right\rfloor\right) + c\left(\left\lceil \frac{1}{2} S_t^{\hat{g}} \right\rceil\right)\right) \\
&= \lim_{t\to\infty} \frac{1}{T} \sum_{t=T_0+1}^{T-1} \left(c\left(\left\lfloor \frac{1}{2} S_t^{\hat{g}} \right\rfloor\right) + c\left(\left\lceil \frac{1}{2} S_t^{\hat{g}} \right\rceil\right)\right) \\
&= \sum_{s=0}^{\infty} \pi^{\hat{g}}(s) \left(c\left(\left\lfloor \frac{1}{2} s \right\rfloor\right) + c\left(\left\lceil \frac{1}{2} s \right\rceil\right)\right) a.s. \qquad (3.93)
\end{aligned}$$

where $\pi^{\hat{g}}(s)$ is the stationary distribution of $\{S_t = S_t^{\hat{g}}, t \geq T_0 + 1\}$. The second equality in (3.93) holds because $T_0 < \infty\, a.s.$; the last equality in (3.93) follows by the Ergodic theorem for irreducible positive recurrent Markov chains [69, chap. 3].
Since the sum $\frac{1}{T}\sum_{t=0}^{T-1} \left(c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right)\right)$ converges $a.s.$, from Corollary III.18 we have

$$\begin{aligned}
J^{\hat{g}}(\pi_0^1, \pi_0^2) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \left(c\left(X^{1,\hat{g}}\right) + c\left(X^{2,\hat{g}}\right)\right) \\
&= \sum_{s=0}^{\infty} \pi^{\hat{g}}(s) \left(c\left(\left\lfloor \frac{1}{2} s \right\rfloor\right) + c\left(\left\lceil \frac{1}{2} s \right\rceil\right)\right). \qquad (3.94)
\end{aligned}$$

Since the right hand side of equation (3.94) does not depend on the initial PMFs π_0^1, π_0^2, we obtain

$$J^{\hat{g}}(\pi_0^1, \pi_0^2) = J^{\hat{g}}(0, 0). \qquad (3.95)$$

Combining (3.95) and Lemma III.22 we get

$$J^{\hat{g}}(\pi_0^1, \pi_0^2) = J^{\hat{g}}(0,0) \leq \inf_{g \in \mathcal{G}_d} J^g(\pi_0^1, \pi_0^2). \tag{3.96}$$

Thus, $\hat{g}$ is an optimal routing policy for the infinite horizon problem.

□

3.7 The Case of Bursty Arrivals

In this section, we consider the same model as that of Section 3.2 except that we allow multiple arrivals to and multiple departures from each queue at each time instant. Suppose the maximum burst of customers at each time is a number K. Then, the arrival process $\{A_t^i, t \in \mathbb{Z}_+\}$ is i.i.d. with $A_t^i \in \{0, 1, 2, \ldots, K\}$ and $\mathbb{E}[A_t^i] = \lambda^i$ for $i = 1, 2$; the departure process $\{D_t^i, t \in \mathbb{Z}_+\}$ is i.i.d. with $D_t^i \in \{0, 1, 2, \ldots, K\}$ and $\mathbb{E}[D_t^i] = \mu^i$ for $i = 1, 2$. The arrival rates λ^1, λ^2 and departure rates μ^1, μ^2 can capture various queueing systems with heterogeneous stations. To ensure the total number of arrivals to the system is less the total number of departures, we make the following assumption.

Assumption III.23. $\lambda^1 + \lambda^2 < \mu^1 + \mu^2$.

In the situation of bursty arrivals, we allow each controller to route up to K customers from its own queue to the other queue at each time. We use $U_t^i \in \{0, 1, 2, \ldots, K\}$ to denote the routing decision of controller C_i at t $(i = 1, 2)$; if $U_t^i = k$, k customers are routed from Q_i to Q_j $(j \neq i)$.

The objective for the system with bursty arrivals is to design decentralized policies that balance the two queues (in the sense we describe below).

3.7.1 The Decentralized Policy $\mathbf{DR}_M$

Given the maximum burst of customer K, we introduce a family of decentralized routing policies $\mathrm{DR}_M := (\mathrm{DR}_M{}^1, \mathrm{DR}_M{}^2)$ for any number $M \geq 1$ as follows:

$$U_t^i = \mathrm{DR}_M{}^i \left(\overline{X}_t^i, \overline{UB}_t, \overline{LB}_t \right) = k, \text{ when } \gamma_t(k) \leq \overline{X}_t^i < \gamma_t(k+1) \tag{3.97}$$

where $\gamma_t(0) = \overline{LB}_t$, and for $k = 1, 2, \ldots, K+1$

$$\gamma_t(k) = \overline{LB}_t + \left\lfloor (2k + M - 1) \max \left(1, \frac{\overline{UB}_t - \overline{LB}_t + 1}{2K + M + 1} \right) \right\rfloor \tag{3.98}$$

where $\lfloor x \rfloor$ denotes the integer part of x.

We will show later in Theorem III.27 that policy DR_M can balance the two queues such that the difference between the queues is at most M for any value of M.

Under DR_M, each controller routes k customers to the other queue when $\overline{X}_t^i, i = 1, 2$, the queue length of its own station at the time of decision, is between the thresholds $\gamma_t(k)$ and $\gamma_t(k+1)$ for $k = 1, 2, \ldots, K$. Note that, the threshold vector γ_t, as a function of $\overline{\Pi}_t^1, \overline{\Pi}_t^2$, is common knowledge between the controllers. Using the common information, each controller can compute the thresholds according to (3.30) individually, and DR_M can be implemented in a decentralized manner.

The evolution of the bounds defined by (3.21)-(3.28) have dynamics described by the following lemma.

Lemma III.24. *At any time t we have*

$$\overline{UB}_t^{DR_M} = UB_t^{DR_M} + K, \quad \overline{LB}_t^{DR_M} = \left(LB_t^{DR_M} - K \right)^+. \tag{3.99}$$

and

$$UB_{t+1}^{i,DR_M} = \gamma_t(U_t^{i,DR_M} + 1) - U_t^{i,DR_M} + U_t^{j,DR_M} - 1 \tag{3.100}$$

$$LB_{t+1}^{i,DR_M} = \gamma_t(U_t^{i,DR_M}) - U_t^{i,DR_M} + U_t^{j,DR_M}. \tag{3.101}$$

Proof. See Appendix B. □

Using the bounds' evolution in Lemma III.4, we obtain the following result.

Lemma III.25. *Under policy DR_M*

$$\begin{aligned}
&UB_{t+1}^{DR_M} - LB_{t+1}^{DR_M} - M \\
&\leq \left\lfloor \frac{2\max(U_t^{1,DR_M}, U_t^{2,DR_M}) + M + 1}{2K + M + 1} \left(\overline{UB}_t^{DR_M} - \overline{LB}_t^{DR_M} - 2K - M \right)^+ \right\rfloor \\
&\leq \left\lfloor \frac{2\max(U_t^{1,DR_M}, U_t^{2,DR_M}) + M + 1}{2K + M + 1} \left(UB_t^{DR_M} - LB_t^{DR_M} - M \right)^+ \right\rfloor \\
&\leq \left(UB_t^{DR_M} - LB_t^{DR_M} - M \right)^+.
\end{aligned} \tag{3.102}$$

Proof. See Appendix B. □

Note that $UB_t - LB_t$ describes the largest difference/gap between the two queue lengths. Lemma III.25 shows that the the difference between the number M and the above largest gap is non-increasing under the policy DR_M.

A direct consequence of Lemma III.25 is the following corollary.

Corollary III.26. *If $UB_{t_0}^{DR_M} - LB_{t_0}^{DR_M} \leq M$ for some time t_0, then*

$$\left(UB_t^{DR_M} - LB_t^{DR_M} \right) \leq M \text{ for all } t \geq t_0. \tag{3.103}$$

Corollary III.26 says that once the difference between the lengths of the two queues becomes less than or equal to M, it will always remain less than or equal to M

afterwards. Therefore, we need to ensure that the event $\{UB_t^{\mathrm{DR}_M} - LB_t^{\mathrm{DR}_M} \leq M\}$ will eventually happen. The following theorem asserts that the above event occurs with probability one.

Theorem III.27. *Let $\tau_M = \min\{t : UB_t - LB_t \leq M\}$ be the first time when the largest difference between the lengths of the two queues is at most M. Then under Assumption III.23, policy DR_M ensures that*

$$\mathbb{E}\left[\tau_M^{DR_M}\right] < \infty. \tag{3.104}$$

Proof. See Appendix B. □

Theorem III.27 guarantees that the two queues will become balanced, so that the largest difference between their lengths is at most M under DR_M. When $M = 1$, the two queues will be eventually balanced (they will be differ by at most one customer); however, the smaller M is, the more customers are routed at each time. When routing cost is taken into account and the total instantaneous cost includes both holding cost and routing cost, DR_M could achieve minimum total cost by selecting an optimal value of M.

3.8 Numerical Example

In this Section, we use numerical simulations to illustrate the results developed in this chapter. Section 3.8.1 presents an example of single arrivals (the model of Section 3.2), and Section 3.8.2 presents an example of bursty arrivals (the model of Section 3.7).

3.8.1 Single Arrivals

We consider the queueing system in Section 3.2 with arrival rate $\lambda = 0.4$ and service rate $\mu = 0.5$. We assume that the holding cost is $c(x) = x^2$, and the initial

PMFs π_0^1 and π_0^2 are both uniformly distributed in the set $[0, 50]$. It is not hard to verify that Assumptions III.10-III.12 are satisfied.

In the numerical simulation, we set the initial queue lengths to be $X_0^1 = 5$ and $X_0^2 = 40$.

Fig. 3.3 shows the evolution over time of the bounds $UB_t^{\hat{g}}, LB_t^{\hat{g}}$ and queue lengths $X_t^{1,\hat{g}}, X_t^{2,\hat{g}}$; it also shows that $\hat{g}$ balances the queue lengths $X_t^{1,\hat{g}}, X_t^{2,\hat{g}}$. Moreover, policy $\hat{g}$ controls the difference between the upper bound $UB_t^{\hat{g}}$ and the lower bound $LB_t^{\hat{g}}$. This difference is non-increasing and converges to one at time $T_0 = 26$ (as asserted in Corollary III.21).

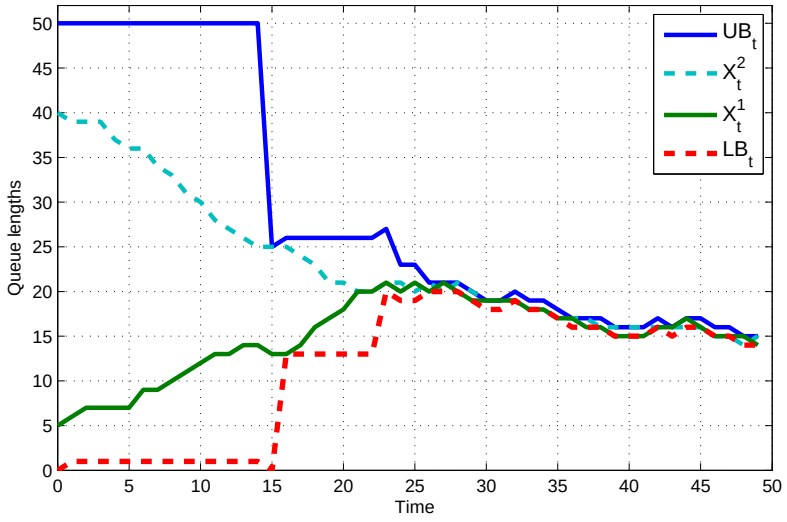

Figure 3.3: The evolution of queue lengths under policy $\hat{g}$

3.8.2 Bursty Arrivals

We consider the queueing system in Section 3.7 with maximum bursts $K = 5$. Let A_t^i be binomial$(5, 0.4)$ and D_t^i be binomial$(5, 0.5)$ for $i = 1, 2$ at each time t. Then $\lambda^1 = \lambda^2 = 2$ and $\mu^1 = \mu^2 = 2.5$. We assume that the initial PMFs π_0^1 and π_0^2 are both uniformly distributed in the set $[0, 50]$.

Since $\lambda^1 + \lambda^2 < \mu^1 + \mu^2$, Assumption III.23 is satisfied.

In the numerical simulation, we set the initial queue lengths to be $X_0^1 = 5$ and $X_0^2 = 40$.

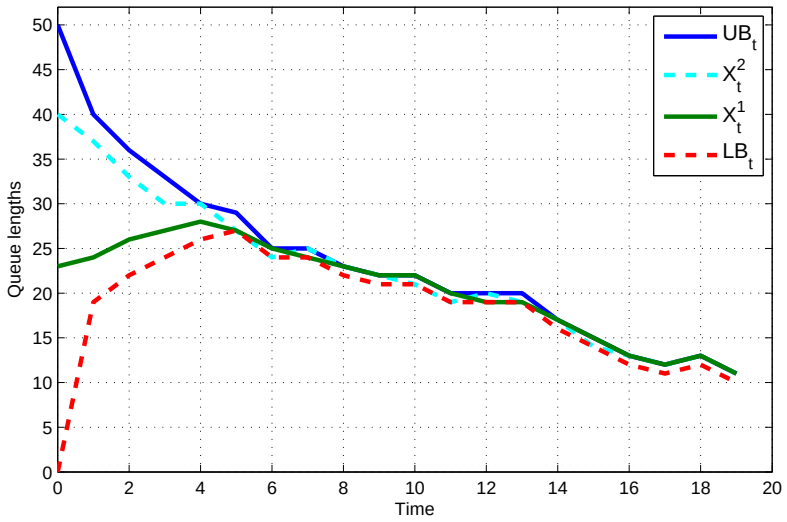

Figure 3.4: The evolution of queue lengths under policy DR_M with $M = 1$

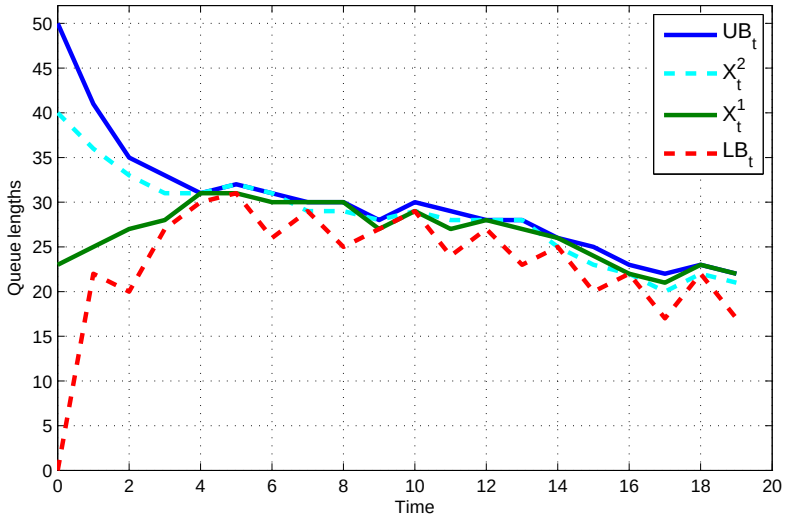

Figure 3.5: The evolution of queue lengths under policy DR_M with $M = 5$

Fig. 3.4 shows the numerical results under DR_M when $M = 1$ and Fig. 3.5 shows the numerical results under DR_M when $M = 5$. In Fig. 3.4, the queue lengths

$X_t^{1,\mathrm{DR}_M}, X_t^{2,\mathrm{DR}_M}$ are eventually balanced by DR_M with $M = 1$. Fig. 3.5 shows that the difference between $X_t^{1,\mathrm{DR}_M}, X_t^{2,\mathrm{DR}_M}$ is eventually less than 5 under DR_M with $M = 5$. These results illustrate the assertion of Theorem III.27.

3.9 Discussion and Conclusion

Based on the results established in Sections 3.3-3.6, we now discuss and answer the questions posed in Section 3.1.

Controllers C_1 and C_2 communicate with one another through their control actions; thus, each controller's information depends on the decision rule/routing policy of the other controller. Therefore, the queueing system considered in this chapter has non-classical information structure [70]. A key feature of the system's information structure is that at each time instant each controller's information consists of one component that is common knowledge between C_1 and C_2 and another component that is its own private information. The presence of common information allows us to use the common information approach, developed in [13], along with specific features of our model to identify an information state/sufficient statistic for the finite and infinite horizon optimization problem. The identification/discovery of an appropriate information state proceeds in two steps: In the first step we use the common information approach (in particular [67]) to identify the general form of an information state (namely $\left(\overline{X}_t^i, \overline{\Pi}_t^1, \overline{\Pi}_t^2\right)$) for controller $C_i, i = 1, 2$ (and the corresponding structure of an optimal policy, Properties III.3). In the second step we take advantage of the features of our system to further refine/simplify the information state; we discover a simpler form of information state, namely, $\left(\overline{X}_t^i, \left\{\overline{UB}_t^j, \overline{LB}_t^j\right\}_{j=1,2}\right)$ for controller $C_i, i = 1, 2$. The component $\left\{\overline{UB}_t^j, \overline{LB}_t^j\right\}_{j=1,2}$ of the above information state describes the common information between controllers C_1 and C_2 at time $t, t = 1, 2, \ldots$.

Using this common information, we established an optimal signaling strategy $\hat{g}$

for the queueing system with signal arrivals, and analyzed the family of signaling strategy DR_M for the system with bursty arrivals.

For the single arrivals case, the update of $\left\{\overline{UB}_t^j, \overline{LB}_t^j\right\}_{j=1,2}$ is described by (3.32)-(3.35) and explicitly depends on the signaling policy $\hat{g}$. Specifically, if a customer is sent from Q_i to Q_j $(i \neq j)$ at time t the lower bound on the queue length of Q_i increases because both controllers know that the length of Q_i is above the threshold TH_t at the time of routing; if no customer is sent from Q_i to Q_j at time t, the upper bound on the length of Q_i decreases because both controllers know that the length of Q_i is below the threshold TH_t at the time of routing. The update of common information incorporates the information about a controller's private information transmitted to the other controller through signaling. Similar updates for the bursty arrivals case are given by (III.24), which depends on the signaling policy DR_M.

The signaling policies $\hat{g}$ and DR_M (with $M = 1$) communicate information in such a way that eventually the difference between the upper bound and the lower bound on the queue lengths is no more than one. Thus, signaling through $\hat{g}$ and DR_M (with $M = 1$) results in a balanced queueing system under single arrivals or bursty arrivals.

CHAPTER IV

Decentralized Stochastic Control-Part II: Multiple Access Communication

4.1 Introduction

Multiple access communication has played a crucial role in the operation of many networked systems, including satellite networks, radio networks, wired/wireless Local Area Networks (LANs), and data centers. One important feature of multiple access communication is its decentralized information structure. In general, when multiple users share the communication system, coordination among them is essential to resolve collision issues. In the absence of a centralized controller, it is challenging to design efficient user coordination mechanisms.

We consider a typical slotted multiple access communication system where multiple users share a common collision channel. Each user is equipped with an infinite size buffer and observes Bernoulli arrivals to its own queue. In addition to the local information, all users receive a common broadcast feedback from the channel. The feedback indicates whether the previous transmission was successful (exactly one user transmitted), or it was a collision (more than one users transmitted), or the channel was idle. The objective is to design a transmission protocol that effectively coordinates the users' transmissions under the above described information structure. In

the design of transmission protocols, there are two major performance metrics of interest: throughput and delay. The throughput region of a protocol is the set of arrival rates for which the users' queues are stable (see detailed definition in Section 4.2.2) under the protocol. The delay performance of a protocol is the average waiting time of a packet in the communication system. An efficient transmission protocol should achieve the maximum throughput region and incur low transmission delay.

In this chapter, we propose a common information (see [13, 71]) based multiple access protocol (CIMA) that uses the common channel feedback to coordinate users. In CIMA, each user constructs upper bounds on the lengths of the queues of all users, including itself, based on previous transmission strategies and the common feedback. Since the upper bounds are common knowledge, users can coordinate their transmissions through these common upper bounds to avoid collision. We prove that without knowledge of any statistics, CIMA achieves the full throughput region of the collision channel. We also prove that the CIMA protocol incurs low transmission delay; the delay is upper-bounded by a linear function of the number of users.

There is rich literature on multiple access communications. Below we present a survey of this literature.

Related Work

There are primarily two classes of protocols for the situation where the alphabet of the feedback channel is $\{0, 1, e\}$ = {no transmission, successful transmission, collision}: collision-free and contention-based protocols. Time-division-multiple-access (TDMA) [72] and adaptive TDMA [73, 74] are collision-free protocols. In adaptive TDMA protocols the (common) information provided by the feedback is used to adaptively coordinate users to avoid collision. Adaptation resolves the problems due to asymmetric arrivals, and collision avoidance results in higher throughput and lower delay than TDMA. However, there is no theoretical analysis of adaptive

TDMA protocols. Backoff-type protocols and Aloha protocols [72] allow for contention/collision. Due to collision, most contention based protocols can not achieve full throughput. However, polynomial back-off protocols, presented and analyzed in [75], achieve full throughput. Nevertheless, polynomial back-off protocols have exponential delay performance in simulation.

Several types of multiple access protocols were proposed when the common information among the users is more than $\{0, 1, e\}$. The authors of [76–78] proposed decentralized random access protocols that achieve full throughput when each user knows the maximum queue length in the system or all other users' transmission results. When channel sensing is allowed, carrier sense multiple access (CSMA) protocols, proposed in [79–84], achieve full throughput when the channel sensing portion of time is not taken into account in the throughput calculation. A survey of CSMA protocols is presented in [85]. In terms of delay performance, the CSMA protocols proposed in [83, 84] achieve delay that is linear in the number of users.

Multiple access protocols for adversarial queueing models were presented in [86, 87]. In [86, 87] it is proved that these protocols achieve full throughput and have linear delay in the number of users.

Other models for multiple access have also been proposed in the literature. In [88], channel switching policies that achieve high throughput for multiple access have been considered within the context of the slotted Aloha protocol and the IEEE 802.11 WLANs protocol. The stability region of the multi-packet reception multiple access channel has been investigated in [89]. Multiple access with noisy channels has been considered in [61, 62], and the stability region of policies with delayed shared information has been determined.

Organization

The rest of the chapter is organized as follows. In Section 4.2 we present the system model and formulate the problem under investigation. In section 4.3 we present the CIMA protocol. In Section 4.4 we prove that the CIMA protocol achieves full throughput and linear (in the number of users) delay. We present simulation results and compare the delay of our protocol with the delay of other protocols that achieve full throughput in Section 4.5. We conclude in Section 4.6. We present proofs of the technical results in Appendix C.

4.2 System Model and Objective

4.2.1 System Model

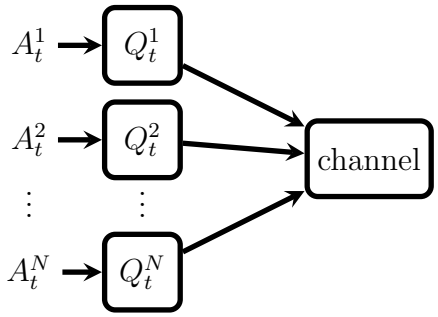

Figure 4.1: Multiple Access Collision Channel.

We consider a slotted communication system, described by Fig. 4.1, where N users, indexed by $1, 2, \ldots, N$, share a common collision channel; we denote by $\mathcal{N} := \{1, 2, \ldots, N\}$ the set of channel users. Each user n is associated with an infinite size buffer with queue length Q_t^n at the beginning of each time slot t. We assume that each queue is initially empty.

At each time slot t each user can transmit one packet in its queue through the shared channel. If only one user transmits in a time slot, the transmission is successful

and the transmitted packet is removed form the queue; if more than one users transmit simultaneously, a collision occurs and all packets involved in the collision remain in their queue. We consider Bernoulli arrivals to the system. Let A_t^n denote the packet arrival to user n at time t; $A_t^n = 1$ means that a packet arrives at queue n right after the transmission at time t. The arrival A_t^n is a Bernoulli random variable with parameter λ^n, and the arrival processes $\{A_t^n, t = 0, 1, \dots\}, n \in \mathcal{N}$ are independent. Let U_t^n denote the transmission decision of user n at time slot t; $U_t^n = 1$ (resp. 0) indicates that user n transmits (resp. does not transmit) at time t. The dynamics of queues are given by

$$Q_{t+1}^n = A_t^n + \left(Q_t^n - U_t^n \prod_{m \neq n} (1 - U_t^m) \right)^+, \tag{4.1}$$

where $(\cdot)^+ := \max(\cdot, 0)$. We assume that at the end of each time slot t, every user receives a feedback $F_t \in \{0, 1, e\}$ from the channel/receiver indicating whether no packets, one packet, or more than one packet (a collision) were transmitted, respectively, in this time slot. This communication system is decentralized; each user can only observe its own queue length, its arrivals and the common feedback. Moreover, the arrival rates $\lambda := (\lambda^1, \lambda^2, \dots, \lambda^N)$ are *not known* to the users. Therefore, the users' decisions according to any decentralized transmission policy/protocol $g = \{g_t^n, n = 1, 2, \dots, N, t = 0, 1, \dots\}$ are generated by

$$U_t^n = g_t^n(Q_{0:t}^n, A_{0:t-1}^n, U_{0:t-1}^n, F_{0:t-1}), \tag{4.2}$$

$n = 1, 2, \dots, N, t = 0, 1, 2, \dots$

In this chapter, we consider throughput and queueing delay as the performance metrics of a decentralized transmission policy/protocol. The objective is to design a decentralized protocol to achieve full throughput and to maintain low queueing delay. We proceed to define the throughput region and queueing delay of the communication

system.

4.2.2 Stability and Throughput Optimality

For queueing systems that can be described by irreducible Markov chains, stability is usually defined to be positive recurrence of the corresponding Markov chains. In this problem, the users' actions can generally depend on the whole history of information. When non-Markovian control policies are used, the resulting queue length processes are not Markov in general. Even within the class of Markovian policies, the corresponding Markov chain may not be irreducible under any Markovian policy.

To achieve higher throughput performance of the communication system, we consider general non-Markovian policies of the form given by (4.2). Therefore, a stability notion for general stochastic processes is essential for our analysis of the system. In this chapter, we call a stochastic process $\{X_t, t = 0, 1 \dots\}$ *stable* if for every $\epsilon > 0$ there exists a finite set K such that

$$\mathbb{P}(X_t \notin K) < \epsilon \text{ for all } t. \tag{4.3}$$

This stability concept is also used in [90–92], and it is called bounded in probability in [93]. Note that the stability criterion (4.3) is equivalent to positive recurrence for countable irreducible Markov chains [93, Proposition 18.3.1]. For general countable Markov chains with a reachable state, bounded in probability is equivalent to positive Harris recurrence, another stability concept for general Markov chains [93, Proposition 18.3.2].

Given the arrival rates $\lambda = (\lambda^1, \dots, \lambda^N)$ to all queues, a policy/protocol g stabilizes the communication system if the resulting queue length process $\{Q_t^{n,g}, t = 0, 1, \dots\}$ for every user $n = 1, \dots, N$ is stable. The arrival rate λ is said to be supportable if there exist policies/protocols that can stabilize the communication system under λ.

For any arrival rates $\lambda = (\lambda^1, \ldots, \lambda^N)$, we use $\lambda^{tot} := \sum_{n=1}^{N} \lambda^n$ to denote the total arrival rate to the communication system. Since at most one packet can be transmitted through the collision channel at each time, only $\lambda \in \Lambda$ could be supportable, where

$$\Lambda = \left\{ \lambda = (\lambda^1, \lambda^2, \ldots, \lambda^N) : \lambda^{tot} < 1 \right\}. \tag{4.4}$$

Furthermore, any $\lambda \in \Lambda$ is supportable by the time sharing policy that assigns λ^n portion of time slots to user n. Therefore, arrival rates λ are supportable if and only if $\lambda \in \Lambda$. We call Λ the throughput region of the multiple access communication system. We call a decentralized policy/protocol throughput optimal if it can stabilize the communication system for any $\lambda \in \Lambda$.

4.2.3 Queueing Delay

Let $Q_t^{tot} := \sum_{n=1}^{N} Q_t^n$ denote total queue length of the system at time $t, t = 1, 2, \ldots$. We define

$$Q_{avg} := \limsup_{t \to \infty} \frac{1}{T} \mathbb{E} \left[\sum_{t=0}^{T-1} Q_t^{tot} \right]. \tag{4.5}$$

From Little's law (see [94]), in a stable queueing system, the queueing delay of a packet is proportional to the average total number of packets in the system. For a throughput optimal protocol g, the queueing delay of the system is given by $\frac{Q_{\text{avg}}^g}{\lambda^{tot}}$.

4.2.4 Objective

Our objective is to find a throughput optimal protocol that results in low queueing delay.

4.3 The Common Information-Based Multiple Access (CIMA) Protocol

4.3.1 Preliminaries

We first introduce common upper bounds for the queues. Let $B_t^g := (B_t^{1,g}, B_t^{2,g}, \ldots, B_t^{N,g})$, where $B_t^{n,g}$ is the upper bound on Q_t^n at time slot t based on the transmission protocol g and the common information $F_{0:t-1}$, received from the common feedback, up to time slot t. That is, when $F_{0:t-1} = f_{0:t-1}$,

$$b_t^{n,g} = \max\{q_t^n : \exists \lambda \in \Lambda \text{ s.t. } \mathbb{P}^{\lambda,g}(q_t^n | f_{0:t-1}) > 0\}.$$

Note that, B_t^g is a function of the common information $F_{0:t-1}$. We use B_t^g to denote that the common upper bounds depend explicitly on the transmission policy g.

4.3.2 The CIMA Protocol

The CIMA protocol is defined as follows.

$$\begin{aligned} U_t^n =& \text{CIMA}_t^n(Q_{0:t}^n, A_{0:t-1}^n, U_{0:t-1}^n, F_{0:t-1}) \\ =& \begin{cases} 1 & \text{if } v(B_t^{\text{CIMA}}) = n \text{ and } Q_t^n > 0, \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{4.6}$$

where $v(\cdot)$ is a function of common upper bounds B_t^{CIMA} defined as

$$v(b_t^{\text{CIMA}}) = \min\{n : b_t^{n,\text{CIMA}} = \max_{m=1,2,\ldots,N} b_t^{m,\text{CIMA}}\}.$$

Note that $v(b_t^{\text{CIMA}})$ is the user with the largest common upper bound. Since we want to avoid collision, if there are more than one users with the largest common upper bound, CIMA selects the user with the smallest index.

4.4 Performance Analysis of the CIMA Protocol

We prove that the CIMA protocol is throughput optimal in 4.4.2. We provide an upper bound on the queueing delay under the CIMA protocol in 4.4.3.

4.4.1 Preliminary Results

In order to analyze the system dynamics under the CIMA protocol, we first provide the following result.

Lemma IV.1. *Under the CIMA protocol, the queue lengths are independent conditional on the common feedback given any arrival rates λ. Specifically, for any time t, any realization $f_{0:t-1}$ and any value $q_t = (q_t^1, \ldots, q_t^N)$ of $Q_t = (Q_t^1, \ldots, Q_t^N)$,*

$$\mathbb{P}^{\lambda,CIMA}(q_t|f_{0:t-1}) = \prod_{n=1}^{N} \mathbb{P}^{\lambda,CIMA}(q_t^n|f_{0:t-1}). \tag{4.7}$$

Moreover, the conditional probability can be updated as follows. For $n \neq v(b_t^{CIMA})$

$$\begin{aligned}&\mathbb{P}^{\lambda,CIMA}(q_{t+1}^n|f_{0:t})\\ =&\lambda^n \mathbb{P}^{\lambda,CIMA}(Q_t^n = q_{t+1}^n - 1|f_{0:t-1}) + (1-\lambda^n)\mathbb{P}^{\lambda,CIMA}(Q_t^n = q_{t+1}^n|f_{0:t-1}).\end{aligned} \tag{4.8}$$

For $n = v(b_t^{CIMA})$ and $f_t = 1$

$$\begin{aligned}&\mathbb{P}^{\lambda,CIMA}(q_{t+1}^n|f_{0:t})\\ =&\lambda^n \frac{\mathbb{P}^{\lambda,CIMA}(Q_t^n = q_{t+1}^n|f_{0:t-1})1_{\{q_{t+1}^n>0\}}}{\mathbb{P}^{\lambda,CIMA}(Q_t^n > 0|f_{0:t-1})} + (1-\lambda^n)\frac{\mathbb{P}^{\lambda,CIMA}(Q_t^n = q_{t+1}^n + 1|f_{0:t-1})}{\mathbb{P}^{\lambda,CIMA}(Q_t^n > 0|f_{0:t-1})}.\end{aligned} \tag{4.9}$$

For $n = v(b_t^{CIMA})$ *and* $f_t = 0$

$$\mathbb{P}^{\lambda,CIMA}(q_{t+1}^n | f_{0:t}) = \begin{cases} 0 & \text{if } q_{t+1}^n \geq 2, \\ \lambda^n & \text{if } q_{t+1}^n = 1, \\ 1 - \lambda^n & \text{if } q_{t+1}^n = 0. \end{cases} \tag{4.10}$$

Proof. See Appendix C. □

Using Lemma IV.1, we can obtain the evolution of queue lengths and common upper bounds under CIMA, as stated in the lemma below.

Lemma IV.2. *Under CIMA, the queue lengths evolve as*

$$Q_{t+1}^{n,CIMA} = \begin{cases} A_t^n + Q_t^{n,CIMA} & \text{if } n \neq v(B_t^{CIMA}), \\ A_t^n + \left(Q_t^{n,CIMA} - 1\right)^+ & \text{if } n = v(B_t^{CIMA}). \end{cases} \tag{4.11}$$

and the common upper bounds evolve according to

$$B_{t+1}^{n,CIMA} = \begin{cases} B_t^{n,CIMA} + 1 & \text{if } n \neq v(B_t^{CIMA}), \\ B_t^{n,CIMA} & \text{if } n = v(B_t^{CIMA}) \text{ and } F_t = 1, \\ 1 & \text{if } n = v(B_t^{CIMA}) \text{ and } F_t = 0. \end{cases} \tag{4.12}$$

Proof. See Appendix C. □

Using Lemma IV.2, the CIMA protocol can be easily implemented as described in Fig 4.2 below.

for $k = 1$ to N **do**

$B^k \leftarrow 0$

end for

while user n is active **do**

$B^{\mathrm{MAX}} \leftarrow \max_k(B^k)$

$v \leftarrow \min(k : B^k = B^{\mathrm{MAX}})$

if $n = v$ and $Q_t^n > 0$ **then**

transmit a packet (set $U_t = 1$)

end if

for $k \neq v$ **do**

$B^k \leftarrow B^k + 1$

end for

if $F_t = 1$ **then**

$B^v \leftarrow B^v$

else

$B^v \leftarrow 1$

end if

end while

Figure 4.2: The CIMA protocol for user $n \in \{1, 2, \ldots, N\}$.

4.4.2 Throughput Optimality

The main result on CIMA's throughput is stated in the following theorem.

Theorem IV.3. *The CIMA protocol is throughput optimal. That is, for any arrival rates $\lambda \in \Lambda$ (defined by (4.4)), the queue length processes under CIMA are stable.*

To prove the theorem, we first show that under the CIMA protocol the queue

lengths together with the upper bounds form a Markov chain.

Lemma IV.4. *Let* $Y_t^{CIMA} := (Q_t^{CIMA}, B_t^{CIMA})$*, where*

$$Q_t^{CIMA} = (Q_t^{1,CIMA}, Q_t^{2,CIMA}, \dots, Q_t^{N,CIMA})$$

for every time slot $t = 0, 1, \dots$ *Then,* $\{Y_t^{CIMA}, t = 0, 1, \dots\}$ *is a Markov chain.*

Proof. See Appendix C. □

Since $\{Y_t^{\text{CIMA}}, t = 0, 1, \dots\}$ is a Markov chain, we can use the Foster-Lyapunov theorem in the proof below to show that the process $\{Y_t^{\text{CIMA}}, t = 0, 1, \dots\}$ is stable.

Proof of Theorem IV.3. Let $\epsilon = 1 - \lambda^{tot}$. Then $\epsilon > 0$ because $\lambda \in \Lambda$. Let $y := (q, b) = (q^1, q^2, \dots, q^N, b^1, b^2, \dots, b^N)$. Define the Lyapunov function $h(y)$ by

$$h(y) = \sum_{n=1}^{N} (q^n + \alpha b^n), \tag{4.13}$$

where $\alpha = \frac{\epsilon}{2(N-1)}$. For $Y_t^{\text{CIMA}} = y$, let $v = v(b) = \min(n : b^n = \max_{k \in \mathcal{N}}(b^k))$. Then from (4.11) and (4.12) in Lemma IV.2 we get

$$\begin{aligned} &\mathbb{E}\left[h(Y_{t+1}^{\text{CIMA}}) - h(Y_t^{\text{CIMA}}) | Y_t^{\text{CIMA}} = y\right] \\ \leq &-\epsilon/2 \quad \text{if } b^v \geq \frac{1}{\alpha} + 1. \end{aligned} \tag{4.14}$$

(see Appendix C for a detailed derivation of (4.14))
Since $b^v = \max_{k \in \mathcal{N}}(b^k)$, $b^v \geq b^n$ and $b^v \geq q^n$ for all $n = 1, 2, \dots, N$. Define

$$C = \{y = (q, b) : q^n < \frac{1}{\alpha} + 1, b^n < \frac{1}{\alpha} + 1 \quad \forall n\}.$$

Then, (4.14) holds for every $y \notin C$. Since C is a finite set, the Foster-Lyapunov drift criterion (Condition (DD2) in [95]) is satisfied. From [95, Theorem 4.5], $\{Y_t^{\text{CIMA}}, t =$

$0, 1, \dots\}$ is bounded in probability (satisfies the stability condition (4.3)).

Therefore, for every $\epsilon > 0$ there exists a finite set K such that

$$\mathbb{P}(Y_t^{\text{CIMA}} \notin K) < \epsilon \text{ for all } t. \tag{4.15}$$

Let $K^n = \{q^n : \text{there exists } y = (q, b) \in K\}$ be the projection of K on its nth component. Then,

$$\mathbb{P}(Q_t^{n,\text{CIMA}} \notin K^n) \leq \mathbb{P}(Y_t^{\text{CIMA}} \notin K) < \epsilon \text{ for all } t. \tag{4.16}$$

Therefore, $\{Q_t^{n,\text{CIMA}}, t = 0, 1, \dots\}$ also satisfies (4.3) and the stability of the communication system under CIMA is established.

□

Remark IV.5. We provide an alternative proof of Theorem IV.3.
As a result of (4.14), condition (V2) in [93, Chap. 11] is satisfied. Therefore, by Theorem 11.3.4 in [93] the Markov chain $\{Y_t^{\text{CIMA}}, t = 0, 1, \dots\}$ is positive Harris recurrent on a countable state space. By Theorem 18.3.2 in [93] positive Harris recurrence implies (4.15), which in turn implies (4.16), and this establishes the assertion of Theorem IV.3.

4.4.3 Delay Performance

Using CIMA, we have the following queueing delay performance guarantee.

Theorem IV.6. *Under the CIMA protocol, for any rate $\lambda \in \Lambda$ we have*

$$\frac{Q_{avg}^{CIMA}}{\lambda^{tot}} \leq \frac{2N}{1 - \lambda^{tot}}. \tag{4.17}$$

Theorem IV.6 says that for any fixed total arrival rate λ^{tot}, the queuing delay under the CIMA protocol is linear in the number of users N.

To prove Theorem IV.6, we first present a property of the CIMA protocol.

Lemma IV.7. *Let* $\bar{U}_t = \sum_{n=1}^{N} U_t^n \prod_{m \neq n}(1 - U_t^m)$. *If the total number of packets at time* t *is* $Q_t^{tot,CIMA} = q$, *there are at least* q *successful transmissions from time* t *to* $t + q + N - 1$ *using the CIMA protocol. That is*

$$\sum_{\tau=t}^{t+q+N-1} \bar{U}_\tau^{CIMA} \geq q.$$

Proof. See Appendix C. □

Since $U_t^n \in \{0, 1\}$ for each $n, n = 1, 2, \ldots, N$, $\bar{U}_t \in \{0, 1\}$; $\bar{U}_t = 1$ (respectively, $\bar{U}_t = 0$) denotes a successful (respectively, unsuccessful) transmission at time t. Lemma IV.7 shows that when at a certain time slot the total queue length is q, the CIMA protocol can successfully transmit at least q packets in the next $q + N - 1$ time slots.

Using Lemma IV.7, we can now prove Theorem IV.6.

Proof of Theorem IV.6. Let $T_1 = N$, and define recursively the random variables $T_2, T_3, \ldots$ by

$$T_k = \left\{ \min t : t > T_{k-1}, \sum_{\tau=T_{k-1}}^{t-1} \bar{U}_\tau^{\text{CIMA}} = Q_{T_{k-1}}^{tot,\text{CIMA}} \right\}.$$

Then, each T_k is the time such that $Q_{T_{k-1}}^{tot,\text{CIMA}}$ packets are successfully transmitted from time T_{k-1} to $T_k - 1$ under the CIMA protocol.

By Lemma IV.7 the CIMA protocol can successfully transmit at least $Q_{T_{k-1}}^{tot,\text{CIMA}}$ packets from time T_{k-1} to $T_{k-1} + Q_{T_{k-1}}^{tot,\text{CIMA}} + N - 1$. Therefore

$$T_k \leq T_{k-1} + Q_{T_{k-1}}^{tot,\text{CIMA}} + N. \tag{4.18}$$

Consequently, from the dynamics of queues and (4.18) we obtain

$$\mathbb{E}\left[Q_{T_k}^{tot,\mathrm{CIMA}}\right] \leq \lambda^{tot}\left(\mathbb{E}\left[Q_{T_{k-1}}^{tot,\mathrm{CIMA}}\right] + N\right). \tag{4.19}$$

(see Appendix C for a detailed derivation of (4.19))

Since $T_1 = N$, $\mathbb{E}\left[Q_{T_1}^{tot,\mathrm{CIMA}}\right] \leq \mathbb{E}\left[\sum_{t=0}^{N-1}\sum_{n=1}^{N} A_t^n\right] = \lambda^{tot}N$. From (4.19), we can show, recursively, that for all k

$$\begin{aligned}\mathbb{E}\left[Q_{T_k}^{tot,\mathrm{CIMA}}\right] \leq & \lambda^{tot}N + (\lambda^{tot})^2 N + \cdots + (\lambda^{tot})^k N \\ \leq & \frac{\lambda^{tot}N}{1-\lambda^{tot}}.\end{aligned} \tag{4.20}$$

Now for any time $t = 0, 1, 2, \ldots$, for any realization of arrivals there is some number k such that $T_{k-1} < t \leq T_k$ ($T_0 := 0$). Using (4.20) and the dynamics of queues we get

$$\mathbb{E}\left[Q_t^{tot,\mathrm{CIMA}}\right] \leq 2\frac{\lambda^{tot}N}{1-\lambda^{tot}} \tag{4.21}$$

(see Appendix C for a detailed derivation of (4.21))

Since (4.21) holds for any time t, we have

$$\begin{aligned}Q_{\mathrm{avg}}^{\mathrm{CIMA}} = & \limsup_{t\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[Q_t^{tot,\mathrm{CIMA}}\right] \\ \leq & \limsup_{t\to\infty} \frac{1}{T}\sum_{t=0}^{T-1}\frac{2\lambda^{tot}N}{1-\lambda^{tot}} \\ = & \frac{2\lambda^{tot}N}{1-\lambda^{tot}}.\end{aligned} \tag{4.22}$$

□

Remark IV.8. The result of Theorem IV.6 implies throughput optimality of the CIMA protocol. Since the bound on delay (the right hand side of (4.17)) is finite for every $\lambda \in \Lambda$, it can be shown that the stability requirement described by (4.3) is satisfied.

Nevertheless, the proof of Theorem IV.3 is interesting/instructive by itself, and for this reason we have proved throughput optimality and delay performance separately.

4.5 Simulation Results

In this section we first compare, via simulation, the queueing delay incurred by CIMA with that of three other protocols that use the same feedback information and no channel sensing: the basic TDMA protocol, the adaptive TDMA (ATDMA) protocol [73] and the quadratic back-off protocol which is proved to be throughput optimal in [75]. In the quadratic back-off protocol, each user transmits a packet with probability $(c+1)^{-2}$ where c is the back-off counter. We also compare the delay performance of CIMA with CSMA protocols proposed in[86, 87]; these protocols employ channel sensing before transmission scheduling.

In the numerical experiments, we have used different values of N and λ^{tot} for each protocol. Arrival rates are asymmetric: half of the users have arrival rate $1.4\lambda^{tot}/N$ and the other half of the users have arrival rate $0.6\lambda^{tot}/N$. For each N and λ^{tot}, we run the simulation for $T = 10^5$ time steps.

The simulation results of Fig. 4.3 show that the average delay associated with the CIMA protocol is linear in the number of users. These simulation results are consistent with the result of Theorem IV.6.

In Fig. 4.4, we compare the delay performance of TDMA, ATDMA, quadratic back-off and CIMA protocols for a system of 4 users. Fig. 4.4 shows that the delay associated with the CIMA protocol is significantly smaller than that of the quadratic back-off protocol (that is also throughput optimal) and of the TDMA protocol (note that TDMA is unstable when $\lambda^{tot} > 0.7$). CIMA's delay is also smaller than the delay of ATDMA (note that there is no theoretical analysis for ATDMA).

In Fig. 4.5, we compare the delay performance of CIMA with two CSMA protocols: the PGD protocol proposed in [83], and the DCSMA protocol proposed in [84]. The

results of [83] and [84] prove that the two CSMA protocols achieve delay that is linear in the number of system users. That is, the delay of the CSMA protocols is of the same order as the delay of the CIMA protocol. However, channel sensing is required to implement the two CSMA protocols. Moreover, Fig. 4.5 shows that the delay resulting from the CIMA protocol is significantly smaller than that of the CSMA protocols of [83] and [84].

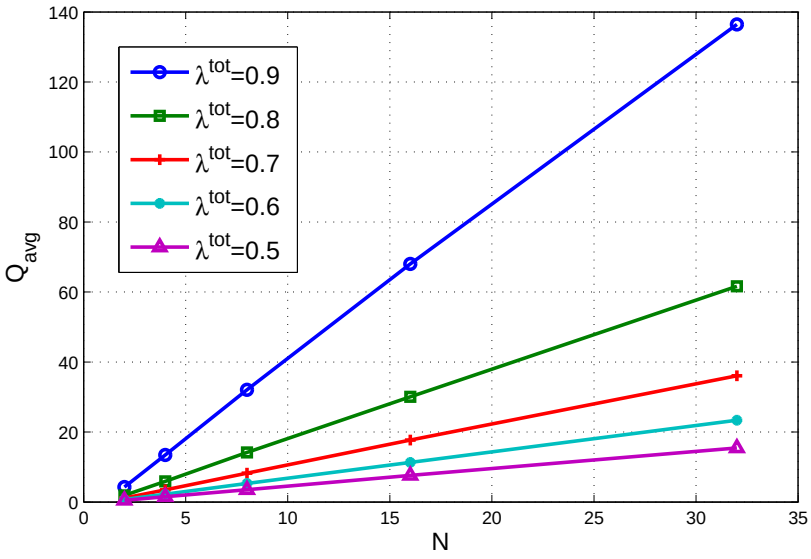

Figure 4.3: Delay versus the number of users of CIMA.

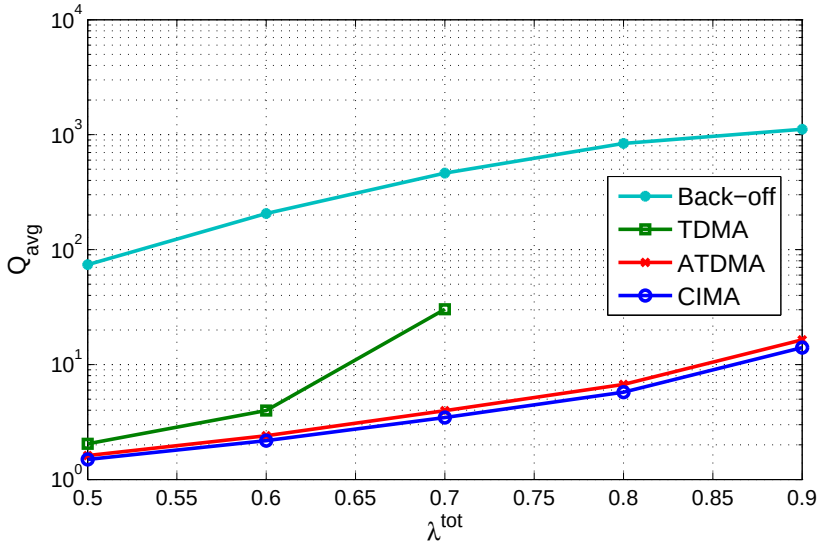

Figure 4.4: Comparison of protocols for a system of 4 users.

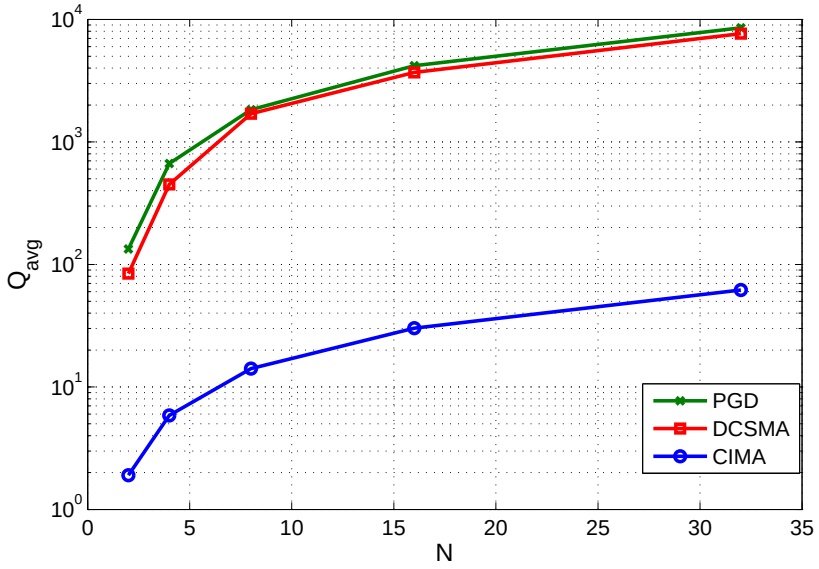

Figure 4.5: Comparison of CIMA and CSMA protocols.

4.6 Conclusion

We developed a transmission protocol that utilizes the common information of the system's users to achieve efficient/optimal coordination of their transmissions. The protocol is collision free; thus, it is similar in spirit to TDMA (or adaptive TDMA), but it differs from TDMA in the way it selects the user to transmit at each time slot. Intuitively, we expect that the delay due to the CIMA protocol will increase linearly with the number of users. The result of Theorem IV.6 confirms this intuition.

The problem investigated in this chapter can be viewed as a decentralized control/decision-making problem with non-classical information structure [70]. Decentralized stochastic control problems with non-classical information structure are signaling problems [3]. In our setup signaling occurs through the feedback provided by the collision channel. Signaling leads to adjustments of each user's upper bounds on their queue lengths (in the manner described by CIMA) and results in efficient coordination among the users. Signaling also occurs in CSMA protocols and in adaptive TDMA, but it is distinctly different from that of the CIMA protocol.

CHAPTER V

Dynamic Stochastic Games with Asymmetric Information

5.1 Introduction

Background and Motivation

Stochastic dynamic games arise in many socio-technological systems such as cyber-security systems, electronic commerce platforms, communication networks, etc. In all these systems, there are many strategic decision makers (agents). In dynamic games with symmetric information all the agents share the same information and each agent makes decisions anticipating other agents' strategies. This class of dynamic games has been extensively studied in the literature (see [5–9] and references therein). An appropriate solution concept for this class of games is sub-game perfect equilibrium (SPE), which consists of a strategy profile of agents that must satisfy *sequential rationality* [5, 6]. When the system state is

The common history in dynamic games with symmetric information can be utilized to provide a sequential decomposition of the dynamic game. The common history (or a function of it) serves as an information state and SPE can be computed through backward induction.

Many instances of stochastic dynamic games involve asymmetric information, that

is, agents have different information over time (such games are also called dynamic games of incomplete information in the game and economic theory literature). In communication networks, different nodes have access to different local observations of the network. In electronic commerce systems, each seller has private information about the quality of his product. In cyber-security systems, a defender cannot directly detect the attacker's activities. In this situation, if an agent wants to assess the performance of any particular strategy, he needs to form beliefs (over time) about the other agents' private information that is relevant to his objective. Therefore, *perfect Bayesian equilibrium* (PBE) is an appropriate solution concept for this class of games. PBE consists of a pair of strategy profile and belief system for all agents that jointly must satisfy *sequential rationality* and *consistency* [5, 6]. In games with asymmetric information a decomposition similar to that of games with symmetric information is not possible in general. This is because the evaluation of an agent's strategy depends, in general, on the agent's beliefs about all other agents' private information over time. Since private information increases with time, the space of beliefs on the agents' private information grows with time. As a result, sequential computation of equilibria for stochastic dynamic games with asymmetric information is available only for special instances (see [14, 17–26] and references therein).

In this chapter, we consider a general model of a dynamic game with a finite number of agents/players in a system with asymmetric information. The information available to an agent at any time can be decomposed into *common information* and *private information*. Common information refers to the part of an agent's information that is known by all agents; private information includes the part of an agent's information that is known only by that agent. We define a class of PBE and provide a sequential decomposition of the game through an appropriate choice of information state using ideas from the common information approach for decentralized decision-making, developed in [13]. The proposed equilibrium and the associated de-

composition resemble Markov perfect equilibrium (MPE), defined in [16] for dynamic games with symmetric information.

Games with asymmetric information have been investigated in the economic literature within the context of repeated games of incomplete information (see [17–20] and references therein). A key feature of these games is the absence of dynamics. The problems investigated in [14, 21–26] are the most closely related to our problem. The authors of [21–25] analyze zero-sum games with asymmetric information. The authors of [14, 26] used a common information based methodology, inspired by [13], to establish the concept of *common information based Markov perfect equilibria*, and to achieve a sequential decomposition of the dynamic game that leads to a backward induction algorithm that determines such equilibria. Our problem is different from those investigated in [14, 21–26] for the following reasons. It is a nonzero-sum game, thus, it is different from the problems analyzed in [21–25]. Our approach to analyzing dynamic games with asymmetric information is similar to that of [14, 26]; the key difference between our problem and those in [14, 26] is in the information structure. The information structure in [14, 26] is such that the agents' common information based (CIB) beliefs are *strategy-independent*, therefore there is *no signaling* effect. This naturally leads to the concept of common information based Markov perfect equilibrium. In our problem the information structure is such that the CIB beliefs are *strategy-dependent*, thus *signaling* is present. In such a case, the specification of a belief system along with a strategy profile is necessary to analyze the dynamic game. Signaling is a key phenomenon present in stochastic dynamic games with asymmetric information. Since it plays a fundamental role in the class of games we investigate in this chapter, we discuss its nature and its role below. The discussion will allow us to clarify the nature of our problem, after we formulate it, and to contrast it with the existing literature, in particular [14, 26].

Signaling

In a dynamic game with asymmetric information, an agent's private information is not observed directly by other agents. Nevertheless, when an agent's strategy depends on his private information, part of this private information may be revealed/transmitted through his actions. We call such a strategy a *private strategy*. When the revealed information from an agent's private strategy is "*relevant*" to other agents, the other agents utilize this information to make future decisions. This phenomenon is referred to as *signaling* in games [4] and in decentralized control [3]. When signaling occurs, agents' beliefs about the system's private information (which is defined to be the union of all agents' private information) depend on the agents' strategies (see [4]). Signaling may occur in games with asymmetric information depending on the system dynamics, the agents' utilities and the information structure of the game. Below we identify game environments where signaling occurs, as well as environments where signaling does not occur.

To identify game environments where signaling occurs we need to precisely define what we mean by the statement: an agent's private information is "*relevant*" to other agents. For that matter we define the concepts of payoff relevant and payoff irrelevant information.

We call a variable (e.g. the system state, an observation, or an action) *payoff relevant* (respectively, *payoff irrelevant*) to an agent at time t if the agent's expected continuation utility at t directly depends on (respectively, does not depend on) this variable given any fixed realization of all other variables[1]. For instance, in a dynamic game with Markov dynamics where agents' utilities at each time only depend on the current states, the current states are payoff relevant and the history of previous states is payoff irrelevant.

[1]Decomposition of agents' types to payoff-relevant type and payoff-irrelevant type is a standard decomposition in the economic literature. Here, we use term 'variable' instead of 'type' to match with the existing literature in control theory. For a more rigorous definition consult with [96, ch.9].

There are four types of game environments depending on the payoff revelance of an agent's private information.

(a) Agent n's private information at t is payoff relevant to him at t and from $t+1$ on, but payoff irrelevant to other agents from $t+1$ on.

In this game environment, agent n may use a private strategy at t because his private information is payoff relevant to him at t. Then, other agents can infer part of agent n's privation information at t based on agent n's action. Although this revealed private information is payoff irrelevant to other agents, they can use it to anticipate agent n's future actions since this information is payoff relevant to agent n's future utility. In this game environment, signaling from agent n to other agents occurs.

(b) Agent n's private information at t is payoff irrelevant to him at t, and is payoff relevant to other agents from $t+1$ on. This class of games includes the classic *cheap-talk game* [97]. In this game environment, other agents form beliefs about agent n's private information at t because it is payoff relevant to them. By using a private strategy, agent n can affect other agents' beliefs about his private information, hence, affect other agents' future decisions. Signaling may occur in this situation if agent n can improve his future utility when he *signals* part of his private information through his actions (e.g. perfectly informative/separating equilibria in the cheap-talk game). There may be no signaling if by revealing part of his private information agent n does not increase his future utility (e.g. uninformative/pooling equilibria in the cheap-talk game).

(c) Agent n's private information at t is payoff relevant to him at t, and payoff relevant to other agents from $t+1$ on. This game environment has both effects discussed in the previous two environments. As a result, we may have signaling or no signaling from agent n, depending on whether or not he can improve his future utility

by using a private strategy. Decentralized team problems are examples where signaling occurs, because signaling strategies can help the collaborating agents to achieve higher utilities (see [98–101] for examples of signaling strategies in decentralized team problems). Pooling equilibria in the classic two-step *signaling game* [4] is an example of no signaling.

(d) Agent n's private information at t is payoff irrelevant to all agents, including himself, from $t+1$ on. In this game environment no signaling occurs. Even if agent n uses a private strategy at t, since his private information is payoff irrelevant from $t+1$ on to all agents, no agent will incorporate it in their future decisions [2]. The model in [14, 26] are examples of this situation where signaling of information does not occur.

When signaling occurs in a game, all agents' beliefs on the system's private information are strategy dependent. Furthermore, each agent's choice of (private) strategy is based on the above mentioned beliefs, as they allow him to evaluate the strategy's performance. This circular dependence between strategies and beliefs makes the computation of equilibria for dynamic games a challenging problem when signaling occurs. This is not the case for games with no signaling effects. In these games, the agents' beliefs are strategy-independent and the circular dependence between strategies and belief breaks. Then, one can directly determine the agents' beliefs first, and then, compute the equilibrium strategies via backward induction [14, 26].

[2] If one of the agents incorporates the belief on this private information in his strategy from $t+1$ on, all other agents may also incorporate it. The argument is similar to situation (b) since all other agents will anticipate about how this agent will act. We note that, agents can use such payoff irrelevant information as a coordination instrument, and therefore, expand their strategy space thereby resulting in additional equilibria. As an example, consider a repeated prisoner's dilemma game with imperfect public monitoring of actions [20, ch. 7]. The agents can form a punishment mechanism that results in new equilibria in addition to the repetition of the stage-game equilibrium. In general, the idea of such a punishment mechanism is used to proof different versions of folk theorem for different setups [20]. However, we do not call this kind of cases signaling because the signals or actions of an agent serves only as a coordination instrument instead of transmitting private information form one agent to other agents.

Organization

The chapter is organized as follows. We introduce the model of dynamic games in Section 5.2. In Section 5.3, we define the solution concept for our model and compare it with that for the standard extensive game form. In Section 5.4, we introduce the concept of CIB-PBE and provide a sequential decomposition of the dynamic game. In Section 5.5, we provide an example of a multiple access broadcast game that illustrates the results of Section 5.4. We prove the existence of CIB-PBE for a subclass of dynamic games in Section 5.6. We conclude in Section 5.7. The proofs of all of our technical results appear in Appendix D.

5.2 System Model

Consider a dynamic game among N strategic agents, indexed by $\mathcal{N} := \{1, 2, \ldots, N\}$, in a system over time horizon $\mathcal{T} := \{1, 2, \cdots, T\}$. Each agent $n \in \mathcal{N}$ is affiliated with a subsystem n. At every time $t \in \mathcal{T}$, the state of the system (C_t, X_t) has two components: $C_t \in \mathcal{C}_t$ denotes the public state, and $X_t := (X_t^1, X_t^2, \ldots, X_t^N) \in \mathcal{X}_t := \mathcal{X}_t^1 \times \mathcal{X}_t^2 \times \cdots \times \mathcal{X}_t^N$, where X_t^n denotes the local state of subsystem $n, n \in \mathcal{N}$. The public state C_t is commonly observed by every agent, and the local state X_t^n is privately observed by agent $n, n \in \mathcal{N}$.

At time t, each agent n simultaneously selects an action $A_t^n \in \mathcal{A}_t^n$. Given the control actions $A_t := (A_t^1, A_t^2, \ldots, A_t^N)$, the public state and local states evolve as

$$C_{t+1} = f_t^c(C_t, A_t, W_t^C), \tag{5.1}$$

$$X_{t+1}^n = f_t^n(X_t^n, A_t, W_t^{n,X}),\ n \in \mathcal{N}, \tag{5.2}$$

where random variables W_t^C and $W_t^{n,X}$ capture the randomness in the evolution of the system, and $C_1, X_1^1, X_1^2, \ldots, X_1^N$ are primitive random variables.

At the end of time t, after the actions are taken, each agent $n \in \mathcal{N}$ observes

$Y_t := (Y_t^1, Y_t^2, \ldots, Y_t^N)$, where

$$Y_t^n = h_t^n(X_t^n, A_t, W_t^{n,Y}) \in \mathcal{Y}_t^n, \tag{5.3}$$

and $W_t^{n,Y}$ denotes the observation noise. From the system dynamics (5.2) and the observations model (5.3), we define, for any $n \in \mathcal{N}, t \in \mathcal{T}$, the probabilities $p_t^n(x_{t+1}^n; x_t^n, a_t)$ and $q_t^n(y_t^n; x_t^n, a_t)$ such that for all $x_{t+1}^n, x_t^n \in \mathcal{X}_t^n$, $y_t^n \in \mathcal{Y}_t^n$ and $a_t \in \mathcal{A}_t := \mathcal{A}_t^1 \times \cdots \times \mathcal{A}_t^N$

$$p_t^n(x_{t+1}^n; x_t^n, a_t) := \mathbb{P}(f_t^n(x_t^n, a_t, W_t^{n,X}) = x_{t+1}^n), \tag{5.4}$$

$$q_t^n(y_t^n; x_t^n, a_t) := \mathbb{P}(h_t^n(x_t^n, a_t, W_t^{n,Y}) = y_t^n). \tag{5.5}$$

We assume that $\mathcal{C}_t, \mathcal{X}_t^n, \mathcal{A}_t^n$ and $\mathcal{Y}_t^n$ are finite sets for all $n \in \mathcal{N}$, for all $t \in \mathcal{T}$.[3] We also assume that the primitive random variables $\{C_1, X_1^n, W_t^C, W_t^{n,X}, W_t^{n,Y}, t \in \mathcal{T}, n \in \mathcal{N}\}$ are mutually independent.

The actions A_t and the observations $Y_t := (Y_t^1, Y_t^2, \ldots, Y_t^N)$ are commonly observed by every agent. Therefore, at time t, all agents have access to the common history H_t^c defined to be

$$H_t^c := \{C_{1:t}, A_{1:t-1}, Y_{1:t-1}\}. \tag{5.6}$$

Including private information, the history H_t^n of agent n's information, $n \in \mathcal{N}$, at t is given by

$$H_t^n := \{X_{1:t}^n, H_t^c\} = \{X_{1:t}^n, C_{1:t}, A_{1:t-1}, Y_{1:t-1}\}. \tag{5.7}$$

[3]The results developed in Section 5.2-5.4 for finite $\mathcal{C}_t, \mathcal{X}_t^n, \mathcal{A}_t^n$ and $\mathcal{Y}_t^n$ still hold when they are continuous sets under some technical assumptions. The results of Section 5.6 require $\mathcal{A}_t^n$ to be finite for all $n \in \mathcal{N}$, $t \in \mathcal{T}$.

Let $\mathcal{H}_t^c$ denote the set of all possible common histories at time $t \in \mathcal{T}$, and $\mathcal{H}_t^n$ denote the set of all possible information histories for agent $n \in \mathcal{N}$ at time $t \in \mathcal{T}$.

Define $H_t := \cup_{n\in\mathcal{N}} H_t^n = \{X_{1:t}, H_t^c\}$ to be the history of states and observations of the whole system up to time t. The hirtory H_t captures the system evolution up to time t.

A behavioral strategy of agent $n, n \in \mathcal{N}$, is defined as a map $g_t^n : \mathcal{H}_t^n \mapsto \Delta(\mathcal{A}_t^n)$ where

$$\mathbb{P}^{g_t^n}(A_t^n = a_t^n | h_t^n) := g_t^n(h_t^n)(a_t^n) \text{ for all } a_t^n \in \mathcal{A}_t^n. \tag{5.8}$$

Let $\mathcal{G}_t^n$ denote the set of all possible behavioral strategies [4] g_t^n of user $n \in \mathcal{N}$ at time $t \in \mathcal{T}$.

At each time $t \in \mathcal{T}$, agent $n, n \in \mathcal{N}$, has a utility

$$U_t^n = \phi_t^n(C_t, X_t, A_t) \tag{5.9}$$

that depends on the state of the system at t, including the public state and all local states, and the actions taken at t by all agents.

Let $g = (g^1, g^2, \ldots, g^N)$ denote the strategy profile of all agents, where $g^n = (g_1^n, g_2^n, \ldots, g_T^n)$. Then, the total expected utility of agent n is given by

$$U^n(g) = \mathbb{E}^g \left[\sum_{t=1}^{T} U_t^n \right] = \mathbb{E}^g \left[\sum_{t=1}^{T} \phi_t^n(C_t, X_t, A_t) \right]. \tag{5.10}$$

Each agent wishes to maximize his total expected utility.

The problem defined above is a stochastic dynamic game with asymmetric information.

[4] The results developed in this chapter also holds when agent n's set of admissible actions depends on his current private state. That is, $A_t^n \in \mathcal{A}_t^n(x_t^n) \subseteq \mathcal{A}_t^n$ and $g_t^n(h_t^n) \in \Delta(\mathcal{A}_t^n(x_t^n))$ when $h_t^n = (x_{1:t}^n, h_t^c)$.

As discussed above, signaling may occur in games of asymmetric information. The game instances that can be captured by our model could belong to any of the four game environments (a)-(d) described in Section 5.1.

5.3 Solution Concept

For non-cooperative static games with complete information (resp. incomplete information), one can use Nash equilibrium (resp. Bayesian Nash equilibrium) as a solution concept. A strategy profile $g=(g^1,\cdots,g^N)$ is a (Bayesian) Nash equilibrium, if there in no agent n that can unilaterally deviate to another strategy g'^n and get a higher expected utility. One can use (Bayesian) Nash equilibrium to analyze dynamic stochastic games. However, the (Bayesian) Nash equilibrium solution concept ignores the dynamic nature of the system and only requires optimality with respect to any unilateral deviation from the equilibrium g at the beginning of the game (time 1). Requiring optimality only against unilateral deviation at time 1 could lead to irrational situations such as non-credible threats [5, 6]. In dynamic games, a desirable equilibrium g should guarantee that there is no profitable unilateral deviation for any agent at any stage of the game. That is, for any $t \in \mathcal{T}$, for any realization $h_t \in \mathcal{H}_t$ of the system evolution, the strategy $g_{t:T} = (g^1_{t:T}, g^2_{t:T}\cdots, g^N_{t:T})$ must be a (Bayesian) Nash equilibrium of the continuation game that follows h_t. This requirement is called *sequential rationality* [5, 6].

In this chapter we study dynamic stochastic games of incomplete asymmetric information. At time t, the system evolution H_t is not completely known to all agents; each agent $n \in \mathcal{N}$ only observes H^n_t and has to form a belief about the complete system evolution H_t up to time t. The belief that agent n forms about H_t depends in general on both H^n_t and $g^{-n}_{1:t}$. Knowing the strategy of the other agents, agent n can make inference about other agents' private information $X^{-n}_{1:t}$ from observing their actions. As pointed out in Section 5.1, this phenomenon is called signaling in games

with asymmetric information. Signaling results in agents' beliefs that depend on the strategy profile g (see the discussion in Section 5.1). Therefore, at an equilibrium such beliefs must be consistent with the equilibrium strategies via Bayes' rule. Moreover, the sequential rationality requirement must be satisfied with respect to the agents' beliefs. We call the collection of all agents' beliefs at all times a *belief system*. A pair of strategy profile and belief system that are mutually sequentially rational and consistent form a *perfect Bayesian equilibrium* (PBE). We use PBE as the solution concept in this chapter to study the dynamic game defined in Sectin 5.2.

We note that the system model we use in this chapter is different from the standard model of extensive game form used in the game theory literatures [5, 6]. Specifically, the model of Section 5.2 is a state space model (that describes the stochastic dynamics of the system), while the extensive game form is based on the intrinsic model [102] whose components are nature's moves and users' actions. The two models are equivalent within the context of sequential dynamic teams [103]. In order to analyze the dynamic game of the state space model of Section 5.2, we need to provide the formal definition of PBE for our model in the following.

5.3.1 Perfect Bayesian Equilibrium

To provide a formal definition of PBE for our state space model defined in Section 5.2, we first define histories of states, beliefs and signaling-free beliefs on histories of states.

Definition V.1 (History of States)**.** The history of states at each time t is defined to be $X_{1:t}$.

Note that the history of states contains the trajectory of local state $X^n_{1:t}$ that is private information of agent $n, n \in \mathcal{N}$.

Definition V.2 (Belief System)**.** Let $\mu^n_t : \mathcal{H}^n_t \mapsto \Delta(\mathcal{X}_{1:t})$. For every history $h^n_t \in \mathcal{H}^n_t$, the map μ^n_t defines a belief for agent $n \in \mathcal{N}$ at time $t \in \mathcal{T}$ which is a probability

distribution on the histories of states $X_{1:t}$. The collection of maps $\mu := \{\mu_t^n, n \in \mathcal{N}, t \in \mathcal{T}\}$ is called a belief system on histories of states.

That is, given a belief system μ, agent $n \in \mathcal{N}$ assigns the probability distribution $\mu_t^n(h_t^n)$ on $\mathcal{X}_{1:t}$ conditioning on the realized history of observations $h_t^n \in \mathcal{H}_t^n$ at $t \in \mathcal{T}$, by

$$\mathbb{P}_\mu(x_{1:t}|h_t^n) := \mu_t^n(h_t^n)(x_{1:t}). \tag{5.11}$$

Then, given the beliefs $\mu_t^n(h_t^n)$ for agent $n \in \mathcal{N}$ at $h_t^n = (x_{1:t}^n, h_t^c) \in \mathcal{H}_t^n$ and a strategy g_t at $t \in \mathcal{T}$, when agent n takes an action $a_t^n \in \mathcal{A}_t^n$, his belief about the system following (h_t^n, a_t^n) is given by $\mathbb{P}_\mu^{g_t}(x_{1:t+1}, y_t, a_t|h_t^n, a_t^n)$ for any $x_{1:t+1} \in \mathcal{X}_{1:t+1}, y_t \in \mathcal{Y}_t, a_t \in \mathcal{A}_t$, where

$$\begin{aligned}&\mathbb{P}_\mu^{g_t}(x_{1:t+1}, y_t, a_t|h_t^n, a_t^n)\\ :=&\mu_t^n(h_t^n)(x_{1:t}) \prod_{k\in\mathcal{N}} p_t^k(x_{t+1}^k; x_t^k, a_t) q_t^k(y_t^k; x_t^k, a_t) \prod_{k\neq n} g_t^k(x_{1:t}^k, h_t^c)(a_t^k).\end{aligned} \tag{5.12}$$

Definition V.3 (Signaling-Free Beliefs)**.** The signaling-free belief system $\hat{\mu} := \{\hat{\mu}_t^n : \mathcal{H}_t^n \mapsto \Delta(\mathcal{X}_{1:t}), n \in \mathcal{N}, t \in \mathcal{T}\}$ is defined on histories of states such that for each $n \in \mathcal{N}$, $t \in \mathcal{T}$, and $h_t^n := (x_{1:t}^n, c_{1:t}, a_{1:t-1}, y_{1:t-1}) \in \mathcal{H}_t^n$

$$\begin{aligned}\hat{\mu}_t^n(h_t^n)(x_{1:t}) :=&\mathbb{P}_{(A_{1:t-1}=a_{1:t-1})}(x_{1:t}|y_{1:t-1}, x_{1:t}^n)\\ &\text{for any } x_{1:t} \in \mathcal{X}_{1:t}.\end{aligned} \tag{5.13}$$

The right hand side of (5.13) gives the conditional probability of $\{X_{1:t} = x_{1:t}\}$ given $\{Y_{1:t-1} = y_{1:t-1}, X_{1:t}^n = x_{1:t}^n\}$ when $A_{1:t-1} = a_{1:t-1}$[5]. This conditional probability is computed using the realization h_t^n of agent n's information, the subsystem dynamics

[5]We can formally define the signaling-free belief by using open loop strategies. Let's denote $g^{a_{1:t-1}}$ to be the open loop strategies where $g^{a_{1:t-1}}(h_\tau^n)(a_\tau^n) = 1$ for any n and any $\tau \leq t-1$. Then the signaling free belief is defined by $\hat{\mu}_t^n(h_t^n)(x_{1:t}) := \mathbb{P}^{g^{a_{1:t-1}}}(x_{1:t}|h_t^n)$.

(5.2), and the observation model (5.3).

Note that the signaling-free belief $\hat{\mu}_t^n(h_t^n)$ is *strategy-independent*. One can think $\hat{\mu}_t^n(h_t^n)$ as the belief generated by the open-loop strategy $(a_1, a_2, \cdots, a_{t-1})$, so there is no signaling and strategy-dependent inference present in the belief system. The role of signaling-free belief will become evident when we talk about consistency in the definition of PBE for the state space model described in Section 5.2.

The beliefs defined above are used by the agents to evaluate the performance of their strategies. Sequential rationality requires that at any time instant, each agent's strategy is his best response under his belief about the system states.

This relation between a strategy profile g and a belief system μ is formally defined as follows.

Definition V.4 (Sequential Rationality)**.** A pair (g, μ) satisfies *sequential rationality* if for every $n \in \mathcal{N}$, $g_{t:T}^n$ is a solution to

$$\sup_{g_{t:T}'^n \in \mathcal{G}_{t:T}^n} \mathbb{E}_\mu^{g_{t:T}'^n, g^{-n}} \left[\sum_{\tau=t}^{T} \phi_\tau^n(C_\tau, X_\tau, A_\tau) | h_t^n \right] \tag{5.14}$$

for every $t \in \mathcal{T}$ and every history $h_t^n \in \mathcal{H}_t^n$, where $\mathbb{E}_\mu^{g_{t:T}'^n, g^{-n}}[\cdot|h_t^n]$ is computed using the probability measure generated from (5.11)-(5.12) using the belief system μ and the strategy profile $(g_{t:T}'^n, g^{-n})$ given the realization h_t^n.

The above definition of sequential rationality does not place any restriction on the belief system. However, rational agents should form their beliefs based on the strategies used by other agents. This consistency requirement is defined as follows.

Definition V.5 (Consistency)**.** A pair (g, μ) satisfies *consistency* if μ can be computed by Bayes' rule whenever possible. That is, for $n \in \mathcal{N}, t \in \mathcal{T}$, such that

$\mathbb{P}^{g_t}_{\mu}(y_t, a_t | h^n_t, a^n_t) > 0$,

$$\mu^n_{t+1}(h^n_{t+1})(x_{1:t+1}) = \mathbf{1}_{\{x^n_{t+1}\}}(h^n_{t+1}) \frac{\mathbb{P}^{g_t}_{\mu}(x_{1:t+1}, y_t, a_t | h^n_t, a^n_t)}{\mathbb{P}^{g_t}_{\mu}(x^n_{t+1}, y_t, a_t | h^n_t, a^n_t)}$$
$$\text{for any } x_{1:t+1} \in \mathcal{X}_{1:t+1} \tag{5.15}$$

where $\mathbb{P}^{g_t}_{\mu}(\cdot | h^n_t, a^n_t)$ is the probability measure given by (5.12). Furthermore, when $\mathbb{P}^{g}_{\mu}(y_t, a_t | h^n_t, a^n_t) = 0$, $\mu^n_{t+1}(h^n_{t+1})$ is a probability distribution in $\Delta(\mathcal{X}_{1:t+1})$ such that

$$\mu^n_{t+1}(h^n_{t+1})(x_{1:t+1}) = 0 \text{ if } \hat{\mu}^n_{t+1}(h^n_{t+1})(x_{1:t+1}) = 0. \tag{5.16}$$

Note that the signaling-free belief system $\hat{\mu}$ is used in (5.16) of the definition for consistency. We will explain in the discussion below the importance of signaling-free beliefs on agents' rational behavior.

Using the above definitions, we define PBE for the stochastic dynamic game with asymmetric information described by the model of Section 5.2.

Definition V.6 (Perfect Bayesian Equilibrium)**.** A pair (g, μ) is called a *perfect Bayesian equilibrium* (PBE) if it satisfies *sequential rationality* and *consistency*.

5.3.2 Discussion

As we mentioned earlier, the state space model and the extensive game form are different but equivalent representations of sequential dynamic teams. We discuss the connection between these two models for dynamic games. The table below summarizes the key components of our state space model and the extensive game form (see [5, 6]).

State Space Model	Extensive Game Form
State X_t	No State
History H_t	History of Actions
History H_t^n	Information Sets
Belief $\mu_t^n(h_t^n)(x_{1:t})$	Belief on an Information Set
Support of $\hat{\mu}_t^n(h_t^n)(x_{1:t})$	Nodes in an Information Set
PBE	PBE

The state variable X_t in the state space model allows us to easily describe the system dynamics by (5.2). Without an explicit state variable in the extensive form, it may be complex to describe and analyze the system dynamics. In the state space model, the system's evolution is captured by the history of states and observations H_t. This is the analogue of the history of (agents' and nature's) actions in the extensive game form that captures the game's evolution trajectory. The history of information H_t^n defined in the state space model includes all information available to agent n at time t. This history determines what agent n knows about the system, and is the analogue of an information set in the extensive game form. Similarly, agent n's belief on histories of states (conditional on H_t^n) in the state space model is the analogue of agent n's belief over an information set in the extensive game form.

Generally, a belief $\mu_t^n(h_t^n)(x_{1:t})$ can have a fixed support that includes the entire state space $\mathcal{X}_{1:t}$. However, a belief on an information set has a variable support that includes nodes in that particular information set. Given a strategy profile, one can determine the belief using the Bayes' rule, given by (5.15), whenever possible for our state space model and (similarly) for the extensive game form model. However, when the denominator is zero in (5.15), or we reach an information set of measure zero in the extensive game form, one needs to assign values for the belief on $\mathcal{X}_{1:t}$ and on the nodes of the information set. In the extensive game form, the consistency condition allows for any arbitrary probability distribution over the nodes of the (reached) information

set of measure zero. However, in our state space model we need to make sure that the belief assigned is consistent with the dynamics of the system. As a result, the belief does not necessarily assign a positive probability to a history of states and must be more carefully defined. This is where signaling-free beliefs play an important role.

To establish the equivalence between the belief $\mu_t^n(h_t^n)(x_{1:t})$ in our state space model and the belief on the information set of the corresponding extensive game form, we introduce the signaling-free beliefs. The signaling-free belief $\mu_t^n(h_t^n)(x_{1:t})$ defined by (5.13) for $h_t^n = (x_{1:t}^n, c_{1:t}, a_{1:t-1}, y_{1:t-1})$ is constructed by actions $A_{1:t-1} = a_{1:t-1}$ conditioned on the history of observations $y_{1:t-1}, x_{1:t}^n$ using the system dynamics. In forming a signaling-free belief no underlying strategy profile is assumed, and we do not make any further inference by tracing back how the observed actions are generated (i.e. the observed actions are generated by an open loop strategy). Therefore, if a history of states $x_{1:t}$ does not belong to the support of the signaling-free belief (i.e. $\hat{\mu}_t^n(h_t^n)(x_{1:t}) = 0$), this history of states $x_{1:t}$ can not happen under any possible strategy profile. A rational agent should not assign positive probability on any history of states that is outside the support of the signaling-free belief. This leads to the second part of the consistency requirement (5.16). With this additional requirement, the definition of consistency in our state space model is the analogue of the consistency in the extensive game form, and the definitions of PBE in the two models become identical.

We note that the signaling-free beliefs are strategy-independent. In systems where any agent's belief on system's states is strategy-independent (e.g. the finite games considered in [14] and linear-Gaussian systems [26]), one can show that for any strategy profile g, the only consistent belief system is the signaling-free belief system $\hat{\mu}$. In this type of systems, consistency is trivially satisfied using the signaling-free belief system $\hat{\mu}$. As a result, it is sufficient to verify sequential rationality to establish a PBE for systems with strategy-independent beliefs.

5.4 Common Information Based Perfect Bayesian Equilibria and Sequential Decomposition

In this section, we introduce the common information based (CIB) belief system and CIB beliefs. The CIB beliefs generally depend on the agents' strategies because of the presence of signaling in dynamic games with asymmetric information. We use CIB beliefs to construct CIB strategy profiles for the agents. Using the concept of CIB belief system and CIB strategy profile, we define a subclass of PBE called *common information based perfect Bayesian equilibria* (CIB-PBE). The main result of this section provides a sequential decomposition for the dynamic game model in Section 5.2; this decomposition leads to a backward induction algorithm to compute CIB-PBE.

5.4.1 Preliminaries

Based on common histories, we first define CIB signaling-free belief system

Definition V.7 (CIB Signaling-Free Belief System)**.** The CIB signaling-free belief system is $\hat{\gamma} := \{\hat{\gamma}_t : \mathcal{H}_t^c \mapsto \Delta(\mathcal{X}_t), t \in \mathcal{T}\}$ where for each $t \in \mathcal{T}$ and $h_t^c = (c_{1:t}, a_{1:t-1}, y_{1:t-1}) \in \mathcal{H}_t^c$, $\hat{\gamma}_t(h_t^c)$ is a belief on states X_t, with

$$\hat{\gamma}_t(h_t^c)(x_t) := \mathbb{P}_{\{A_{1:t-1}=a_{1:t-1}\}}(x_t|y_{1:t-1}) \text{ for } x_t \in \mathcal{X}_t. \tag{5.17}$$

The right hand side of (5.17) is interpreted in the same way as the right hand side of (5.13)[6]. Note that, $\hat{\gamma}_t(h_t^c)(x_t^{-n}) = \hat{\mu}_t^n(h_t^n)(x_t^{-n})$ from its definition, when $h_t^n = (x_{1:t}^n, h_t^c)$ for any $n \in \mathcal{N}$. We use

$$\hat{\Pi}_t := \hat{\gamma}_t(H_t^c) \tag{5.18}$$

[6]Similar to (5.13), we can formally define the CIB signaling-free belief by $\hat{\gamma}_t^n(h_t^c)(x_t) := \mathbb{P}^{g^{a_{1:t-1}}}(x_t|h_t^c)$, where $g^{a_{1:t-1}}$ denotes the open loop strategies.

to denote the CIB signaling-free belief at time t. Similar to signaling-free beliefs on histories of states defined in (5.13), the CIB signaling-free belief $\hat{\Pi}_t$ depends only on the system dynamics and the observation model.

The CIB signaling-free beliefs have the following dynamics.

Lemma V.8 (Evolution of CIB Signaling-Free Beliefs)**.** *The CIB signaling-free beliefs* $\{\hat{\Pi}_t, t \in \mathcal{T}\}$ *can be updated by*

$$\hat{\Pi}_{t+1} = \prod_{n=1}^{N} \hat{\Pi}^n_{t+1}, \text{ where} \tag{5.19}$$

$$\hat{\Pi}^n_{t+1} = \hat{\psi}^n_t(Y^n_t, A_t, \hat{\Pi}^n_t), \tag{5.20}$$

$$\begin{aligned}&\hat{\psi}^n_t(y^n_t, a_t, \hat{\pi}^n_t)(x^n_{t+1})\\ :=&\frac{\sum_{x^n_t \in \mathcal{X}^n_t} p^n_t(x^n_{t+1}; x^n_t, a_t) q^n_t(y^n_t; x^n_t, a_t) \hat{\pi}^n_t(x^n_t)}{\sum_{x'^n_t \in \mathcal{X}^n_t} q^n_t(y^n_t; x'^n_t, a_t) \hat{\pi}^n_t(x'^n_t)}.\end{aligned} \tag{5.21}$$

Proof. See Appendix D. □

Similar to the belief system defined in Section 5.2, we need a belief system to form an equilibrium. We define CIB belief systems based on the agents' common histories together with CIB update rules.

Definition V.9 (CIB Belief System and CIB Update Rule)**.** A collection of maps $\gamma := \{\gamma_t : \mathcal{H}^c_t \mapsto \Delta(\mathcal{X}_t), t \in \mathcal{T}\}$ is called a CIB belief system. A set of belief update functions $\psi = \{\psi^n_t : \mathcal{Y}^n_t \times \mathcal{A}_t \times \mathcal{C}_t \times \Delta(\mathcal{X}_t) \times \Delta(\mathcal{X}_t) \mapsto \Delta(\mathcal{X}^n_t), n \in \mathcal{N}, t \in \mathcal{T}\}$ is called a CIB update rule.

From any CIB update rule ψ, we can construct a CIB belief system γ_ψ by the following inductive construction:

1. $\gamma_{\psi,1}(h^c_1)(x_1) := \mathbb{P}(x_1) = \prod_{n \in \mathcal{N}} \mathbb{P}(x^n_1) \quad \forall x_1 \in \mathcal{X}_1$.

2. At time $t+1$, after $\gamma_{\psi,t}(h_t^c)$ is defined, set

$$\begin{aligned}&\gamma_{\psi,t+1}(h_{t+1}^c)(x_{t+1}^n)\\ :=&\psi_t^n(y_t^n, a_t, c_t, \gamma_{\psi,t}(h_t^c), \hat{\gamma}_t(h_t^c))(x_{t+1}^n),\end{aligned} \tag{5.22}$$

$$\gamma_{\psi,t+1}(h_{t+1}^c)(x_{t+1}) := \prod_{n=1}^{N} \gamma_{\psi,t+1}(h_{t+1}^c)(x_{t+1}^n), \tag{5.23}$$

for every history $h_{t+1}^c = (h_t^c, c_{t+1}, a_t, y_t) \in \mathcal{H}_{t+1}^c$ and for all $x_{t+1} \in \mathcal{X}_{t+1}$.

For a CIB belief system γ_ψ, we use $\Pi_t^{\gamma_\psi}$ to denote the belief, under γ_ψ, on X_t conditional on H_t^c; that is,

$$\Pi_t^{\gamma_\psi} := \gamma_{\psi,t}(H_t^c) \in \Delta(\mathcal{X}_t). \tag{5.24}$$

We also define the marginal beliefs on X_t^n at time t as

$$\Pi_t^{n,\gamma_\psi}(x_t^n) := \gamma_{\psi,t}(H_t^c)(x_t^n) \quad \forall x_t^n \in \mathcal{X}_t^n. \tag{5.25}$$

Since the CIB beliefs $\{\Pi_t^{\gamma_\psi}, t \in \mathcal{T}\}$ are common information to all agents, all agents can use $\Pi_t^{\gamma_\psi}$ to evaluate the performance of their strategies. Furthermore, if a CIB update rule ψ is properly chosen, the CIB signaling-free belief $\hat{\Pi}_t$ and the CIB belief $\Pi_t^{\gamma_\psi}$ together can summarize the agents' common knowledge about the current system states X_t from all previous actions $A_{1:t-1}$ and observations $Y_{1:t-1}$ available to all of them at time t. This motivates the concept of CIB strategies defined below.

Definition V.10 (CIB Strategy Profile)**.** We call a set of functions $\lambda = \{\lambda_t^n : \mathcal{X}_t^n \times \mathcal{C}_t \times \Delta(\mathcal{X}_t) \times \Delta(\mathcal{X}_t) \mapsto \Delta(\mathcal{A}_t^n), n \in \mathcal{N}, t \in \mathcal{T}\}$ a CIB strategy profile.

For notational simplicity, let $\mathcal{B}_t := \mathcal{C}_t \times \Delta(\mathcal{X}_t) \times \Delta(\mathcal{X}_t)$ and

$$b_t = (c_t, \pi_t, \hat{\pi}_t) \in \mathcal{B}_t \tag{5.26}$$

denote the realization of the part of common information used in a CIB strategy.

If agent n uses a CIB strategy λ_t^n, then any action $a_t^n \in \mathcal{A}_t^n$ is taken by agent n at time t with probability $\lambda_t^n(x_t^n, b_t)(a_t^n)$ when $X_t^n = x_t^n \in \mathcal{X}_t^n$ $(C_t, \Pi_t^{\gamma_\psi}, \hat{\Pi}_t^{\gamma_\psi}) = b_t \in \mathcal{B}_t$. Note that the domain $\mathcal{X}_t^n \times \mathcal{B}_t$ of a CIB strategy λ_t^n is different from the domain $\mathcal{H}_t^n$ of a behavioral strategy g_t^n. However, given a CIB strategy profile λ and a CIB update rule ψ, we can construct a behavioral strategy profile $g \in \mathcal{G}$ by

$$g_t^n(h_t^n) := \lambda_t^n(x_t^n, c_t, \gamma_{\psi,t}(h_t^c), \hat{\gamma}_t(h_t^c)). \tag{5.27}$$

In the following we provide a definition of a CIB belief system consistent with a CIB strategy profile.

Definition V.11 (Consistency)**.** For a given CIB strategy λ_t^n of user $n \in \mathcal{N}$ at $t \in \mathcal{T}$, we call a belief update function ψ_t^n consistent with λ_t^n if (5.28) below is satisfied when the denominator of (5.28) is non-zero;

$$\begin{aligned} &\psi_t^n(y_t^n, a_t, b_t)(x_{t+1}^n) \\ =&\frac{\sum_{x_t^n \in \mathcal{X}_t^n} p_t^n(x_{t+1}^n; x_t^n, a_t)\eta_t^n(x_t^n, y_t^n, a_t, b_t)\pi_t^n(x_t^n)}{\sum_{x_t'^n \in \mathcal{X}_t^n} \eta_t^n(x_t'^n, y_t^n, a_t, b_t)\pi_t^n(x_t'^n)}, \end{aligned} \tag{5.28}$$

where

$$\eta_t^n(x_t^n, y_t^n, a_t, b_t) := q_t^n(y_t^n; x_t^n, a_t)\lambda_t^n(x_t^n, b_t)(a_t^n). \tag{5.29}$$

When the denominator of (5.28) is zero,

$$\psi_t^n(b_t, a_t, y_t^n)(x_{t+1}^n) = 0 \text{ if } \hat{\psi}_t^n(\hat{\pi}_t, a_t, y_t^n)(x_{t+1}^n) = 0. \tag{5.30}$$

For any $t \in \mathcal{T}$, if ψ_t^n is consistent with λ_t^n for all $n \in \mathcal{N}$, we call ψ_t consistent with λ_t. If ψ_t is consistent with λ_t for all $t \in \mathcal{T}$, we call the CIB update rule $\psi = (\psi_1, \ldots, \psi_T)$ consistent with the CIB strategy profile $\lambda = (\lambda_1, \ldots, \lambda_T)$.

Remark V.12. Note that when the denominator of (5.28) is zero, $\psi_t^n(b_t, a_t, y_t^n)$ can be arbitrarily defined as a probability distribution in $\Delta(\mathcal{X}_{t+1})$ satisfying (5.30) and consistency still holds. One simple choice is to set $\psi_t^n(y_t^n, a_t, b_t) = \hat{\psi}_t^n(y_t^n, a_t, \hat{\pi}_t)$ when the denominator of (5.28) is zero; this choice trivially satisfies (5.30). Thus, for any CIB strategy profile λ, there always exists at least a CIB update rule that is consistent with λ.

The following lemma establishes the relation between the consistency conditions given by Definition V.5 and V.11.

Lemma V.13. *If λ is a CIB strategy profile along with its consistent CIB update rule ψ, there exists a pair, denoted by $(g, \mu) = f(\lambda, \psi)$, such that g is the strategy profile constructed by* (5.27) *from (λ, ψ), and μ is a belief system consistent with the strategy profile g. Furthermore, for all $h_t^n \in \mathcal{H}_t^n$, $x_{1:t} \in \mathcal{X}_{1:t}$*

$$\mu_t^n(h_t^n)(x_{1:t}) = \mathbf{1}_{\{x_{1:t}^n\}}(h_t^n) \prod_{k \neq n} \mu_t^c(h_t^c)(x_{1:t}^k) \tag{5.31}$$

where $\mu_t^c : \mathcal{H}_t^c \mapsto \Delta(\mathcal{X}_{1:t})$ satisfies the relation

$$\mu_t^c(h_t^c)(x_t^k) = \sum_{x_{1:t}^{-k} \in \mathcal{X}_{1:t}^{-k}, x_{1:t-1}^k \in \mathcal{X}_{1:t-1}^k} \mu_t^c(h_t^c)(x_{1:t}) = \gamma_{\psi,t}(h_t^c)(x_t^k) \tag{5.32}$$

for all $h_t^c \in \mathcal{H}_t^c$, $k \in \mathcal{N}$ and $x_t^k \in \mathcal{X}_t^k$.

Proof. See Appendix D. □

Lemma V.13 implies that using a CIB strategy profile λ along with its consistent update rule ψ we can construct a behavioral strategy profile g along with its consistent belief system μ.

Note that, equation (5.31) in Lemma V.13 implies that for any agent n, his local states $X_{1:t}^n$ are independent of $X_{1:t}^{-n}$ under μ conditional on any history $h_t^n \in \mathcal{H}_t^n$. Furthermore, the conditional independence described by (5.31) still holds even when agent n uses another strategy since the right hand side of (5.31) depends only on the CIB update rule ψ. This fact is made precise in the following lemma.

Lemma V.14 (Conditional Independence)**.** *Suppose λ is a CIB strategy profile and ψ is a CIB update rule consistent with λ. Let $(g, \mu) = f(\lambda, \psi)$. If every agent $k \neq n$ uses the strategy g^k along with the belief system μ, then under any policy g'^n of agent n, agent n's belief about the states $X_{1:t}$ for $h_t^n \in \mathcal{H}_t^n$ is given by*

$$\begin{aligned}&\mathbb{P}^{g'^n, g^{-n}}(x_{1:t}|h_t^n) = \mu_t^n(h_t^n)(x_{1:t})\\ =&\mathbf{1}_{\{x_{1:t}^n\}}(h_t^n) \prod_{k\neq n} \mu_t^c(h_t^c)(x_{1:t}^k) \text{ for all } x_{1:t} \in \mathcal{X}_{1:t}.\end{aligned} \tag{5.33}$$

Proof. See Appendix D. □

According to Lemma V.14, for any $(g, \mu) = f(\lambda, \psi)$ generated from Lemma V.13, we use

$$\mathbb{P}_\mu(x_{1:t}^{-n}|h_t^c) := \mu_t^n(h_t^n)(x_{1:t}^{-n}) = \mu_t^c(h_t^c)(x_{1:t}^{-n}) \tag{5.34}$$

to indicate that $\mu_t^n(h_t^n)(x_{1:t}^{-n})$ depends only on h_t^c and μ.

5.4.2 Common Information Based Perfect Bayesian Equilibria

Based on the concept of CIB beliefs and CIB strategies, we focus on CIB-PBE defined below.

Definition V.15 (CIB-PBE)**.** A pair (λ^*, ψ^*) of a CIB strategy profile λ^* and a CIB update rule ψ^* is called a Common Information Based Perfect Bayesian Equilibrium (CIB-PBE) if ψ^* is consistent with λ^* and the pair $(g^*, \mu^*) = f(\lambda^*, \psi^*)$ defined in Lemma V.13 forms a PBE.

The following lemma plays a crucial role in establishing the main results of this chapter.

Lemma V.16 (Closeness of CIB Strategies)**.** *Suppose λ is a CIB strategy profile and ψ is a CIB update rule consistent with λ. If every agent $k \neq n$ uses the CIB strategy λ^k along with the belief generated by ψ, then, there exists a CIB strategy λ'^n that is a best response for agent n under the belief generated by ψ at every history $h_t^n \in \mathcal{H}_t^n$ for all $t \in \mathcal{T}$.*

Proof. See Appendix D. □

Lemma V.16 says that the set of CIB strategies is closed under the best response mapping. Since sequential rationality (Definition V.4) requires a strategy profile to be a fixed point under the best response mapping (see (5.14)), Lemma V.16 allows us to restrict attention to the set of CIB strategies to find a fixed point and to search for CIB-PBE.

Below we provide a sequential decomposition of the dynamic game of Section 5.2 that enables us to sequentially compute CIB-PBE via dynamic programming.

In order to sequentially compute CIB-PBE, we define a stage game for each time $t \in \mathcal{T}$ as follows.

Definition V.17. (Stage Game G_t) Given a set of functions $V_{t+1} = \{V_{t+1}^n : \mathcal{X}_t^n \times \mathcal{B}_t \mapsto \mathbb{R}, n \in \mathcal{N}\}$ and a belief update function ψ_t, for any realization $b_t = (c_t, \pi_t, \hat{\pi}_t) \in \mathcal{B}_t$ we define the following Bayesian game $G_t(V_{t+1}, \psi_t, b_t)$.

Stage Game $G_t(V_{t+1}, \psi_t, b_t)$

- There are N players indexed by $\mathcal{N}$.
- Each player $n \in \mathcal{N}$ observes private information $X_t^n \in \mathcal{X}_t^n$; $b_t = (c_t, \pi_t, \hat{\pi}_t)$ are common information.
- $X_t = (X_t^1, \ldots, X_t^N)$ has a prior distribution π_t.
- Each player $n \in \mathcal{N}$ selects an action $A_t^n \in \mathcal{A}_t^n$.
- Each player $n \in \mathcal{N}$ has utility

$$\begin{aligned} & U^n_{G_t(V_{t+1},\psi_t,b_t)} \\ :=& \phi_t^n(c_t, X_t, A_t) + V_{t+1}^n(X_{t+1}^n, B_{t+1}), \text{ where} \qquad (5.35) \\ & B_{t+1} := (C_{t+1}, \psi_t(Y_t, A_t, b_t), \hat{\psi}_t(Y_t, A_t, \hat{\pi}_t)). \qquad (5.36) \end{aligned}$$

If for each $t \in \mathcal{T}$, the functions V_{t+1} are associated with the agents' future utilities, the Bayesian game $G_t(V_{t+1}, \psi_t, b_t)$ becomes a stage game at t of the original game defined in Section 5.2. Therefore, we consider Bayesian Nash equilibria (BNE) of the game $G_t(V_{t+1}, \psi_t, b_t)$. For all V_{t+1} and ψ_t, we define, the BNE correspondence as follows

Definition V.18. (BNE Correspondence)

$$\begin{aligned} & BNE_t(V_{t+1}, \psi_t) := \\ & \{\lambda_t : \forall b_t \in \mathcal{B}_t, \lambda_t|_{b_t} \text{ is a BNE of } G_t(V_{t+1}, \psi_t, b_t), \\ & \quad \text{where } \lambda_t^n|_{b_t}(x_t^n) := \lambda_t^n(x_t^n, b_t) \, \forall \, n \in \mathcal{N}, x_t^n \in \mathcal{X}_t^n\}. \qquad (5.37) \end{aligned}$$

If $\lambda_t|_{b_t}$ is a BNE of $G_t(V_{t+1}, \psi_t, b_t)$, then for all $n \in \mathcal{N}$ and any realization $x_t^n \in \mathcal{X}_t^n$, any $a_t^n \in \mathcal{A}_t^n$ such that $\lambda_t^n|_{b_t}(x_t^n)(a_t^n) > 0$ should satisfy

$$a_t^n \in \arg\max_{a_t'^n \in \mathcal{A}_t^n} \left\{ \mathbb{E}_{\pi_t}^{\lambda_t^{-n}} \left[U^n_{G_t(V_{t+1},\psi_t,b_t)} | x_t^n, a_t'^n \right] \right\}. \tag{5.38}$$

Similar to the dynamic program in stochastic control, for each time $t \in \mathcal{T}$ we define the value update function $D_t^n(V_{t+1}, \lambda_t, \psi_t)$ for each $n \in \mathcal{N}$.

Definition V.19. (Value Update Function)

$$D_t^n(V_{t+1}, \lambda_t, \psi_t)(x_t^n, b_t) := \mathbb{E}_{\pi_t}^{\lambda_t} \left[U^n_{G_t(V_{t+1},\psi_t,b_t)} | x_t^n \right]. \tag{5.39}$$

If $V_t^n = D_t^n(V_{t+1}, \lambda_t, \psi_t)$, for any realization $x_t^n \in \mathcal{X}_t^n$ and $b_t \in \mathcal{B}_t$, the value $V_t^n(x_t^n, b_t)$ denotes player n's expected utility under the strategy profile $\lambda_t|_{b_t}$ in game $G_t(V_{t+1}, \psi_t, b_t)$.

Using the concept of stage games and value update functions, we provide a dynamic programming method to sequentially compute CIB-PBE in the following theorem.

Theorem V.20. *(Sequential Decomposition) A pair (λ^*, ψ^*) of a CIB strategy profile λ^* and a CIB update rule ψ^* is a CIB-PBE if (λ^*, ψ^*) solves the dynamic program for the value functions $V_t^n(\cdot), n \in \mathcal{N}, t \in \mathcal{T} \cup \{T+1\}$ defined by (5.40)-(5.43) below.*

$$V_{T+1}^n(\cdot) := 0 \quad \forall n \in \mathcal{N}; \tag{5.40}$$

for all $t \in \mathcal{T}$

$$\lambda_t^* \in BNE_t(V_{t+1}, \psi_t^*), \tag{5.41}$$

$$\psi_t^* \textit{ is consistent with } \lambda_t^*, \tag{5.42}$$

$$V_t^n = D_t^n(V_{t+1}, \lambda_t^*, \psi_t^*) \quad \forall n \in \mathcal{N}. \tag{5.43}$$

Proof. See Appendix D. □

Note that, from the dynamic program, the value function $V_1^n(x_1^n, (c_1, \pi_1, \pi_1))$ at time $t = 1$ gives agent n's expected utility corresponding to the CIB-PBE (λ^*, ψ^*) conditional on his private information $X_1^n = x_1^n$ and public state $C_1 = c_1$ when the prior distribution of X_1 is π_1.

Using Theorem V.20, we can compute CIB-PBE of the dynamic game. The following algorithm uses backward induction to compute CIB-PBE based on Theorem V.20.

1: $V_{T+1} \leftarrow 0$, $\lambda^* \leftarrow \emptyset$, $\psi^* \leftarrow \emptyset$

2: **for** $t = T$ to 1 **do**

3: **for** every $b_t \in \mathcal{B}_t$ **do**

4: Construct the stage game $G_t(V_{t+1}, \psi_t, b_t)$

5: Compute $(\lambda_t^*|_{b_t}, \psi_t^*|_{b_t})$ such that $\psi_t^*|_{b_t}$ is consistent with $\lambda_t^*|_{b_t}$, and $\lambda_t^*|_{b_t}$ is a BNE of $G_t(V_{t+1}, \psi_t^*|_{b_t}, b_t)$

6: **for** every $n \in \mathcal{N}$ **do**

7: $\lambda_t^{*n}(x_t^n, b_t) \leftarrow \lambda_t^*|_{b_t}(x_t^n)$, $x_t^n \in \mathcal{X}_t^n$

8: $\psi_t^{*n}(y_t, a_t, b_t) \leftarrow \psi_t^*|_{b_t}(y_t, a_t)$, $(y_t, a_t) \in \mathcal{Y}_t \times \mathcal{A}_t$

9: $V_t^n(x_t^n, b_t) \leftarrow D_t^n(V_{t+1}, \lambda_t^*, \psi_t^*)(x_t^n, b_t)$, $x_t^n \in \mathcal{X}_t^n$

10: **end for**

11: **end for**

12: $\lambda^* \leftarrow (\lambda_t^*, \lambda^*)$, $\psi^* \leftarrow (\psi_t^*, \psi^*)$

13: **end for**

Figure 5.1: Backward Induction for Computing CIB-PBE.

Note that in line 5, for different $(\lambda_t^*|_{b_t}, \psi_t^*|_{b_t})$ the algorithm will produce different CIB-PBE. Finding the pair $(\lambda_t^*|_{b_t}, \psi_t^*|_{b_t})$ in line 5 of Fig. 5.1 requires solving a fixed point problem to get a BNE along with a consistent belief system. The complexity for this step is the same as the complexity of finding a PBE for a two-stage dynamic game.

5.5 Example: Multiple Access Broadcast Game

In this section, we illustrate the sequential decomposition developed in Section 5.4 with an example of a two-agent multiple access broadcast system.

Consider a multiple access broadcast game where two agents, indexed by $\mathcal{N} = \{1, 2\}$, share a common collision channel over time horizon $\mathcal{T}$. At time t, $W_t^n \in \{0, 1\}$ packets arrive at each agent $n \in \mathcal{N}$ according to independent Bernoulli processes with $\mathbb{P}(W_t^1 = 1) = \mathbb{P}(W_t^2 = 1) = p = 0.5$. Each agent can only store one packet in his local buffer/queue. Let $X_t^n \in \mathcal{X}_t^n = \{0, 1\}$ denote the queue length (number of packets) of agent n at the beginning of t. If a packet arrives at agent n when his queue is empty, the packet is stored in agent n's buffer; otherwise, the packet is dropped, and agent n incurs a dropping cost of $c = 2$ units.

At each time t, agent n can transmit $A_t^n \in \mathcal{A}_t^n = \{0, 1\}$ packets through the shared channel. If only one agent transmits, the transmission is successful and the transmitted packet is removed from the queue. If both agents transmit simultaneously, a collision occurs and both collided packets remain in their queues. We assume that any packet arriving at time t, $t \in \mathcal{T}$, can be transmitted after t. Then, the queue length processes have the following dynamics. For $n = 1, 2$

$$X_{t+1}^n = \min\left\{X_t^n - A_t^n(1 - A_t^{-n}) + W_t^n, 1\right\}. \tag{5.44}$$

Assume that agents' transmission results at t are broadcast at the end of time t. Then agent n's transmission decision A_t^n at time t is made based on his history of observation $H_t^n = (X_{1:t}^n, A_{1:t-1})$ that consists of his local queue lengths and all previous transmissions from both agents.

Suppose each agent gets a unit reward at t if there is a successful transmission at t. Then, agent n's utility at time t is the reward minus the (expected) dropping cost given by

$$\begin{aligned} &U_t^n = \phi_t^n(X_t, A_t) = \\ &A_t^n \oplus A_t^{-n} - c\,\mathbb{P}(X_t^n - A_t^n(1 - A_t^{-n}) + W_t^n > 1 | X_t, A_t) \end{aligned} \tag{5.45}$$

where $x \oplus y$ denotes the binary XOR operator, and $n \in \mathcal{N}$.

The multiple access broadcast game described above is an instance of the general dynamic model described in Section 5.2. In the following, we use Algorithm 5.1 developed in Section 5.4 to compute a CIB-PBE of this multiple access broadcast game for two time periods, i.e. $\mathcal{T} = \{1, 2\}$.

Before applying Algorithm 5.1, we note some special features of this multiple access broadcast game. First, there is no C_t, Y_t in this multiple access broadcast game. Second, since the private state X_t^n can take only values in $\mathcal{X}_t^n = \{0, 1\}$, any CIB belief in $\Delta(\mathcal{X}_t^n)$ can be described by a number $\pi_t^n \in [0, 1]$ for all $n = 1, 2, t = 1, 2$. Furthermore, given any realization $b_t = (\pi_t, \hat{\pi}_t) \in \mathcal{B}_t$, any CIB strategy $\lambda_t^n(x_t^n, b_t), x_t^n \in \{0, 1\}$, of agent n can be characterized by a number $\beta_t^n \in [0, 1]$ where

$$\beta_t^n := \lambda_t^n(1, b_t)(1). \tag{5.46}$$

This is because A_t^n is binary, and $\lambda_t^n(0, b_t)(1) = 0$ because no packet can be transmitted from an empty queue.

We now use Algorithm 5.1 to sequentially compute a CIB-PBE of the multiple access broadcast game.

Construction of the stage game at $t = 2$

At $t = 2$, for any $b_2 = (\pi_2, \hat{\pi}_2) \in \mathcal{B}_2$, we construct the stage game $G_2(b_2)$ which is a Bayesian finite game (no need to consider a CIB update function because this is the last stage).

Computation of BNE at $t = 2$

Using standard techniques for static games, we obtain a BNE of $G_2(b_2)$ that is

characterized by $\beta_2^*(b_2) = (\beta_2^{*1}(b_2), \beta_2^{*2}(b_2))$, and $\beta_2^*(b_2)$ is given by

$$\beta_2^*(b_2) = \begin{cases} (1,1) & \text{if } \pi_2^1, \pi_2^2 < c^*, \\ (0,1) & \text{if } \pi_2^1 < c^*, \pi_2^2 \geq c^*, \\ (1,0) & \text{if } \pi_2^1 \geq c^*, \pi_2^2 < c^*, \\ (\frac{c^*}{\pi_2^1}, \frac{c^*}{\pi_2^2}) & \text{if } \pi_2^1, \pi_2^2 \geq c^*, \end{cases} \tag{5.47}$$

where $c^* := \frac{1+cp}{2+cp}$. Then we obtain a CIB strategy $\lambda_2^{*n}(1, b_2)(1) = \beta_2^{*n}(b_2)$ for $n = 1, 2$ at time $t = 2$.

Value functions' update at $t = 2$

$V_2^n(x_2^n, b_2) = D_2^n(\lambda_2^*)(x_2^n, b_2), n = 1, 2$, are given by

$$V_2^n(1, b_2) = \begin{cases} 1 - \pi_2^{-n}(1 + cp) & \text{if } \pi_2^1, \pi_2^2 < c^*, \\ \pi_2^{-n} - cp & \text{if } \pi_2^n < c^*, \pi_2^{-n} \geq c^*, \\ 1 & \text{if } \pi_2^n \geq c^*, \pi_2^{-n} < c^*, \\ c^* - cp & \text{if } \pi_2^1, \pi_2^2 \geq c^*. \end{cases} \tag{5.48}$$

$$V_2^n(0, b_2) = \begin{cases} \pi_2^{-n} & \text{if } \pi_2^1, \pi_2^2 < c^*, \\ \pi_2^{-n} & \text{if } \pi_2^n < c^*, \pi_2^{-n} \geq c^*, \\ 0 & \text{if } \pi_2^n \geq c^*, \pi_2^{-n} < c^*, \\ c^* & \text{if } \pi_2^1, \pi_2^2 \geq c^*. \end{cases} \tag{5.49}$$

Construction of the stage game at $t = 1$

At $t = 1$, for any $b_1 = (\pi_1, \hat{\pi}_1) \in \mathcal{B}_1$ and a CIB update function ψ_1, we construct the stage game $G_1(V_2, \psi_1, b_1)$ such that each player $n, n = 1, 2$, has utility

$$\begin{aligned} & U^n_{G_1(V_2, \psi_1, b_1)} \\ =& \phi_1^n(X_1, A_1) + V_2^n(X_2^n, (\psi_1(A_1, b_1)), \hat{\psi}_1(A_1, \hat{\pi}_1))). \end{aligned} \tag{5.50}$$

Then, when the players use CIB strategies λ_1 characterized by $\beta_1 = (\beta_t^1, \beta_1^2)$, from the system dynamics (5.44) and the agents' utilities (5.45), player 1's expected utilities conditional on $X_t^n = 0, 1$, are given by

$$
\begin{aligned}
&\mathbb{E}^{\lambda_1}_{\pi_1}[U^1_{G_1(V_2,\psi_1,b_1)}|X_t^1=1] \\
=&\pi_1^2\beta_1^2 - cp + \beta_1^1(1+cp-\pi_1^2\beta_1^2(2+cp)) \\
&+ \beta_1^1\pi_1^2\beta_1^2\, V_2^n(1,\psi_1((1,1),b_1),\hat{\psi}_1((1,1)),\hat{\pi}_1) \\
&+ \beta_1^1(1-\pi_1^2\beta_1^2)[p\, V_2^n(1,\psi_1((1,0),b_t),\hat{\psi}_1((1,0),\hat{\pi}_1)) \\
&+ (1-p)\, V_2^n(0,\psi_1((1,0),b_t),\hat{\psi}_1((1,0),\hat{\pi}_1))] \\
&+ (1-\beta_1^1)\pi_1^2\beta_1^2\, V_2^n(1,\psi_1((0,1),b_t),\hat{\psi}_1(\hat{\pi}_1,(0,1))) \\
&+ (1-\beta_1^1)(1-\pi_1^2\beta_1^2)\, V_2^n(1,\psi_1((0,0),b_t),\hat{\psi}_1(\hat{\pi}_1,(0,0))), \qquad (5.51) \\
&\mathbb{E}^{\lambda_1}_{\pi_1}[U^1_{G_1(V_2,\psi_1,b_t)}|X_t^1=0] \\
=&\pi_1^2\beta_1^2 + \pi_1^2\beta_1^2[p\, V_2^n(1,\psi_1((0,1),b_t),\hat{\psi}_1((0,1),\hat{\pi}_1)) \\
&+ (1-p)\, V_2^n(0,\psi_1((0,1),b_t),\hat{\psi}_1((0,1),\hat{\pi}_1))] \\
&+ (1-\pi_1^2\beta_1^2)[p\, V_2^n(1,\psi_1((0,0),b_t),\hat{\psi}_1((0,0),\hat{\pi}_1)) \\
&+ (1-p)\, V_2^n(0,\psi_1((0,0),b_t),\hat{\psi}_1((0,0),\hat{\pi}_1))]. \qquad (5.52)
\end{aligned}
$$

Player 2's expected utilities are given by similar equations.

Computation of BNE and belief update function at $t = 1$

When the players use CIB strategies λ_1 characterized by $\beta_1 = (\beta_t^1, \beta_1^2)$, from (5.28) and (5.30), we obtain a CIB update function ψ_1, given below, that is consistent with λ_1 (we select $\psi_1^n(a_1, b_1) = \hat{\psi}_1^n(a_1, \hat{\pi}_1)$ when the denominator of (5.28) is zero).

$$
\psi_1^n(a_1,b_1) = \begin{cases} 1 & \text{if } a_1^n = 1, a_1^{-n} = 1, \\ p & \text{if } a_1^n = 1, a_1^{-n} = 0, \\ \frac{p+\pi_1^n(1-p-\beta_1^n)}{1-\pi_1^n\beta_1^n} & \text{if } a_1^n = 0. \end{cases} \qquad (5.53)
$$

Substituting (5.53) into (5.51) and (5.52), we have the utilities of the two players in game $G_1(V_2, \psi_1, b_1)$. We numerically compute a BNE of $G_1(V_2, \psi_1^*, b_1)$, characterized by $\beta_1^*(b_1) = (\beta_1^{*1}(\pi_1), \beta_1^{*2}(\pi_1))$ such that ψ_1^* satisfies (5.53) when $\beta_1 = \beta_1^*$. The values of $\beta_1^{*1}(\pi_1)$ and $\beta_1^{*2}(\pi_1)$ are shown in Fig. 5.2 for different $\pi_1 \in \Delta(\mathcal{X}_t) = [0,1] \times [0,1]$.

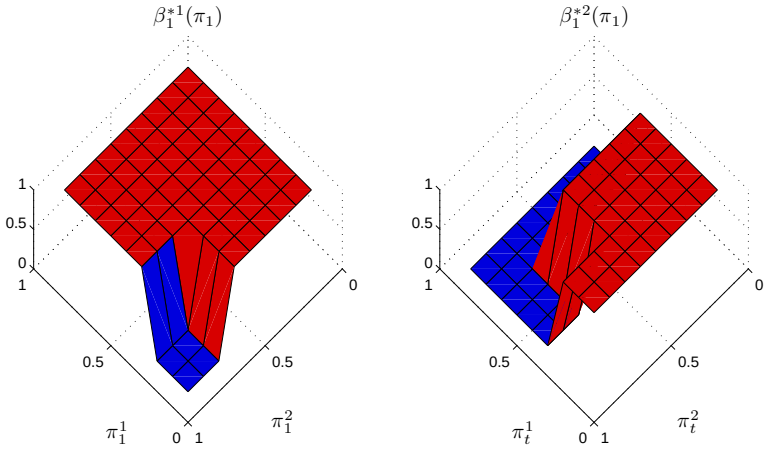

Figure 5.2: Strategies $\beta_1^{*1}(\pi_1)$ and $\beta_1^{*2}(\pi_1)$ in the stage game at time $t = 1$.

Then, we obtain a CIB strategy $\lambda_1^{*n}(1, b_1)(1) = \beta_1^{*n}(\pi_1)$ for $n = 1, 2$ at time $t = 1$.

CIB-PBE and agents' expected utilities

From the above computation at $t = 1, 2$, we obtain a CIB-PBE (λ^*, ψ^*), where $\lambda^* = (\lambda_1^*, \lambda_2^*)$ and $\psi^* = (\psi_1^*, -)$.

Using (5.51) and (5.52), we numerically compute the value functions $V_1^1(x_1^1, b_1) = V_1^1(x_1^1, \pi_1) = D_1^1(V_2, \lambda_1^*, \psi_1^*)(x_1^1, b_1)$ for $x_t^1 = 0, 1$, for agent 1. The results are shown in Fig. 5.3. These value functions give agent 1's (conditional) expected utilities in the CIB-PBE (λ^*, ψ^*). Agent 2's expected utilities can be computed in a similar way.

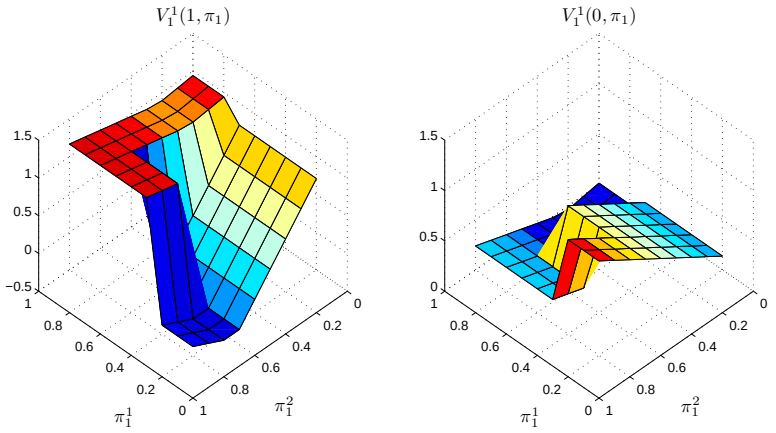

Figure 5.3: Agent 1's expected utility $V_1^n(x_1^n, \pi_1)$ in the CIB-PBE (λ^*, ψ^*).

Remark V.21. The results show that agents' beliefs depend on their strategies (see (5.53)). Therefore, there is signaling in this multiple access broadcast game. Moreover, the value functions are discontinuous in the agents' beliefs (see (5.48) and (5.49) for time $t = 2$, and Fig. 5.3 for time $t = 1$). The presence of signaling together with the discontinuity of value functions make the agents' utilities discontinuous in their (behavioral) strategies.

5.6 Existence of Common Information Based Perfect Bayesian Equilibria

We prove the existence of a CIB-PBE for a subclass of the dynamic games described in section 5.2. This subclass includes dynamic games with uncontrolled dynamics and no private values. No private values simply means that each agent's private information X_t^n is payoff irrelevant to himself, but possibly payoff relevant to the other agents. The classic cheap-talk game [5] and the problem considered in [104] are examples of this subclass. We conjecture that there always exists a CIB-PBE for the general model described in Section 5.2. We discuss this conjecture and elaborate more on the difficulty of establishing an existence proof for the general model of

Section 5.2 at the end of this section.

To proceed formally, let **Game M** denote a dynamic game with uncontrolled dynamics, no private values, finite action spaces $\mathcal{A}_t^n$, $n \in \mathcal{N}$, $t \in \mathcal{T}$, and (possibly) sequential moves. Let $\overline{\mathcal{T}} := \{t_1, t_2, \cdots, t_K\} \subset \mathcal{T}$ denote the set of time instants in which the system evolves according to the following uncontrolled dynamics

$$X_{t+1}^n = \begin{cases} X_t^n & \text{if } t \neq t_k \text{ for all } t_k \in \overline{\mathcal{T}}, \\ f_{t_k}^n(X_{t_k}^n, W_{t_k}^{n,X}) & \text{if } t = t_k \text{ for } t_k \in \overline{\mathcal{T}}. \end{cases} \tag{5.54}$$

At $t_k < t \leq t_{k+1}$ agents make decisions sequentially in $t_{k+1} - t_k$ epochs. We assume that the order according to which the agents take decisions is known a priori. Furthermore, agents observe the other agents' decisions in previous epochs; this fact is captured by including/appending previous actions in the common state C_t as follows

$$C_{t+1} = \begin{cases} (C_t, A_t) & \text{if } t \neq t_k \text{ for all } t_k \in \overline{\mathcal{T}}, \\ f_{t_k}^c(C_{t_{k-1}+1}, W_{t_k}^C) & \text{if } t = t_k \text{ for } t_k \in \overline{\mathcal{T}}. \end{cases} \tag{5.55}$$

The agents have a common observation Y_{t_k} at each time $t_k \in \overline{\mathcal{T}}$ when the system evolves. The observations $Y_t^n, n \in \mathcal{N}, t \in \mathcal{T}$ are described by

$$Y_t^n = \begin{cases} \text{empty} & \text{if } t \neq t_k \text{ for all } t_k \in \overline{\mathcal{T}}, \\ h_{t_k}^n(X_{t_k}^n, W_{t_k}^{n,Y}) & \text{if } t = t_k \text{ for } t_k \in \overline{\mathcal{T}}. \end{cases} \tag{5.56}$$

Agent n, $n \in \mathcal{N}$ has instantaneous utility

$$U_t^n = \phi_t^n(C_t, X_t^{-n}, A_t). \tag{5.57}$$

for time t, $t \in \mathcal{T}$. Thus, each agent $n \in \mathcal{N}$ has no private values, hence his private information X_t^n is payoff irrelevant.

From the above description, it is evident that **Game M** is indeed a subclass of the class of dynamic games described by the model of Section 5.2. The dynamic oligopoly game presented in [104] is an instance of **Game M**.

The main result of this section is stated in the theorem below and asserts the existence of a CIB-PBE in **Game M**.

Theorem V.22. ***Game M*** *described in this section has a CIB-PBE which is a solution to the dynamic program defined by* (5.40)-(5.43) *in Theorem V.20.*

Proof. See Appendix D. □

The proof of Theorem V.22 is constructive. We construct an equilibrium for **Game M** in which agents use non-private strategies and have signaling-free beliefs which are consistent with the non-private strategy profile.

There are three reasons why **Game M** has a CIB-PBE with non-private strategies. First, the instantaneous utility $U_t^n = \phi_t^n(C_t, X_t^{-n}, A_t)$ of agent $n, n \in \mathcal{N}$ does not depend on his private information. Therefore, the agent's best response is the same for all realizations of his private information, and a private strategy doest not provide any advantage in terms of higher instantaneous utility. Second, the system dynamics are strategy independent. Therefore, an agent cannot affect the evolution of the system by using a strategy that depends on his private information about the state of the system. Third, any private strategy does not provide any advantage to an agent in terms of his utility if it can not affect other agents' decisions, and this is the case when all agents use the signaling-free beliefs.

As we showed before, the CIB-PBE introduced in this chapter are PBE. It is known that for finite dynamic games there always exists one sequential equilibrium, and therefore one PBE [5, 6]. The proof of existence of sequential equilibria is indirect; it is done by showing the existence of a trembling hand equilibrium [5, 6] which is also a sequential equilibrium. The proof of existence of trembling hand equilibrium

follows the standard argument in game theory. It uses a suitable fixed point theorem to show the existence of a trembling equilibrium in an equivalent agent-based model representation [5, 6].

There are some technical difficulties in establishing the existence of a CIB-PBE for the general game model considered in this chapter. The standard argument in using fixed point theorems is applicable to finite games where the expected utilities are continuous in the agent's mixed strategies. In each stage game arising in the sequential decomposition, say the game at stage t, agent n's expected utility (see (5.35)) depends on the functions $\{V_{t+1}^n, n \in \mathcal{N}\}$. However, the function V_{t+1}^n is not always continuous in the strategies of agent n. (see Remark V.21 for the multiple access broadcast game in Section 5.5 and the example in [104]). Therefore, the standard argument for establishing the existence of an equilibrium fails for our general model. Even though we can not prove the existence of a CIB-PBE equilibrium, we conjecture that there always exists a CIB-PBE for the general dynamic game described in this chapter.

We note that for the problem formulated in Section 5.2, $\{X_t^t, C_t, \Pi_t, \hat{\Pi}_t\}$, $t \in \mathcal{T}$, are sufficient statistics from the decision making point of view (i.e. control theory). This makes a CIB strategy a more natural strategy choice for an agent, and consequently, a CIB-PBE is a more plausible equilibrium to arise in practice. However, this does not imply that from the game theory point of view, at all equilibria agents' best responses can be generated using only $\{X_t^t, C_t, \Pi_t, \hat{\Pi}_t\}$. In a game problem, agents can incorporate a payoff irrelevant information in their strategy choice as a coordination instrument. For example, consider the classic repeated prisoner's dilemma game. In this game, agents can use previous outcomes of the game, that are payoff irrelevant, to sustain a punishment mechanism that results in additional equilibria beyond the repetition of the stage-game equilibrium [20]. The indirect proof for existence of sequential equilibria and PBE (described above) allows for this type of equilibria that depend on payoff irrelevant information for coordination. Nevertheless, we con-

jecture that there always exists an equilibrium for the game described in Section 5.2 that depends only on $\{X_t^t, C_t, \Pi_t, \hat{\Pi}_t\}$. The example of the dynamic multiple access broadcast game in Section 5.5 is an instance of a dynamic game that does not belong to the subclass of **Game M**, but has a CIB-PBE. To make our conjecture more precise, we provide below a sufficient condition for the existence of a CIB-PBE.

Let (g^*, μ^*) be a strategy profile that is a PBE. Consider the following condition.

Condition C: For all $h_t^c, h_t'^c \in \mathcal{H}_t^c$ such that

$$\mathbb{P}_{\mu^*}(x_t|h_t^c) = \mathbb{P}_{\mu^*}(x_t|h_t'^c), \forall x_t \in \mathcal{X}_t, \forall n \in \mathcal{N}, \forall t \in \mathcal{T}, \tag{5.58}$$

we have

$$g_t^{n*}(x_{1:t}^n, h_t^c) = g_t^{n*}(x_{1:t}^n, h_t^{c'}), \tag{5.59}$$

for all $x_{1:t}^n \in \mathcal{X}_{1:t}^n$.

If **Condition C** is satisfied for a PBE (g^*, μ^*), then a CIB-PBE exists. We conjecture that **Condition C** is satisfied for at least a PBE in the dynamic games considered in this chapter.

5.7 Conclusion

We studied a general class of stochastic dynamic games with asymmetric information. We identified game environments that can lead to signaling in dynamic games. We considered a state space model to analyze dynamic games with private states and controlled Markovian dynamics. We provided a comparison between our state space model and classic extensive game form model. We showed that the two models are equivalent as long as one ensures that the belief system associated with the state space model is compatible with the system dynamics. To ensure the compatibility, we

introduced the signaling-free belief system. Using the signaling-free belief system, we provided a formal definition of PBE in our state space model. We used the common information among agents to define a subset of PBE called CIB-PBE that consist of a pair of CIB strategy profile and CIB update rule. We obtained a sequential decomposition for dynamic games that leads to a backward induction algorithm for the computation of CIB-PBE even when signaling occurs. We illustrated our results with an example of multiple access broadcast game. We proved the existence of CIB-PBE for a subclass of dynamic games and provided a sufficient condition for the existence of CIB-PBE for the general class of dynamic games considered in this chapter.

CHAPTER VI

Conclusion and Future Directions

6.1 Summary

Information interacts with decisions dynamically in modern networked systems. Understanding the interaction between information and decisions is crucial to the analysis of system operations and to the design of efficient decision-making strategies. In this thesis, we study the information-decision interaction in dynamic networked systems under different information structures and different models on DMs' behavior. In all these dynamic systems, the space of available information increases with time since each DM dynamically collects information over time. The analysis of the interaction between information and decisions is a complicated problem due to the increasing space of information, and the fact that each DM's strategy is a mapping from his available information to possible actions. We resolve the difficulty associated with the increasing space of information by identifying information states with fixed domain for the DMs. The nature of information states suggests decision strategies with a corresponding structure. We summarize below the information states considered in this thesis within the context of centralized stochastic control, decentralized stochastic control, and dynamic stochastic games with asymmetric information.

Channel Ordering and PMF Approximation

For centralized stochastic control, we investigated a channel sensing problem in Chapter II, and identified two information states for the problem: the channel ordering and the probability mass functions (PMF) approximation. The channel ordering is an information state that requires an ordering of the channels. We discovered conditions under which the channels can always be ordered. Then the channel ordering allowed us to focus on ordering based policies, and to show the optimality of the myopic policy for the channel sensing problem. The PMF approximation is another information state for channel sensing. We used the PMF approximation to develop an approximation scheme for channel sensing and constructed a near optimal policy. The two information states are complementary. It is more restrictive to use channel ordering as compared to the PMF approximation, but the channel ordering guarantees optimality while the PMF approximation only provides near-optimal results.

Common Belief's Support

For decentralized stochastic control, we investigated the signaling effect between information and decisions in decentralized routing (Chapter III) as well as in multiple access communication (Chapter IV). In both problems, the DMs form common beliefs about the current system states based on their common information. Instead of using the entire complex common beliefs, we focused on their supports and used them as information states for decentralized decision-making. These supports take integer values and have relatively simpler evolution. We designed explicit signaling strategies for the DMs to transmit/signal information through their common beliefs' supports. In both problems, signaling through the common supports allows the DMs to coordinate their decisions and efficiently control the dynamic networked systems.

Common Information Based (CIB) Belief

For dynamic stochastic games with asymmetric information, in Chapter V we introduced a subclass of perfect Bayesian equilibria (PBE) called common information based perfect Bayesian equilibria (CIB-PBE). A CIB-PBE consists of a CIB strategy profile and a CIB update rule. Using the CIB update rule, the DMs construct CIB beliefs that are consistent with their CIB strategy profile. The CIB beliefs serve as information states for the DMs in the dynamic game; all signaling possibilities among the DMs are captured by the evolution of the CIB beliefs through the CIB update rule. Using the above property of CIB beliefs, we developed a sequential decomposition for dynamic games with signaling. The decomposition leads to a backward induction algorithm to compute CIB-PBE.

6.2 Future Directions

As mentioned earlier, using appropriate information states one can reduce the complexity of the interaction between information and decisions in dynamic networked systems. This thesis provides some useful information states for specific problems within the context of centralized stochastic control, decentralized stochastic control, and dynamic stochastic games with asymmetric information. However, the determination/identification of appropriate information states for systems with general information structure and strategic or non-strategic DMs remains an open problem. We present some thoughts on the possible future directions that may deepen our understanding of the information-decision interaction and enhance our ability to identify appropriate information states and to design systems that operate efficiently.

For dynamic networked systems with non-strategic DMs, our goal is to design strategies that are optimal with respect to some pre-specified performance metric. From the common information approach, we know that the common belief on the

system state and all private information is an information state sufficient for performance evaluation. Therefore, strategies using the entire belief can achieve optimal performance. But searching over the space of all possible beliefs is a high complexity problem. One direction to reduce the complexity is to discover properties for reachable beliefs. When the set of reachable beliefs has certain properties, we can reduce the complexity of the optimization problem by searching over strategies that are based only on reachable beliefs. The channel sensing problem studied in Chapter III is an example where any reachable belief satisfies the property that channels can be ordered. Similar ordering properties may hold in other networked systems with imperfectly observed Markovian dynamics. Considering stability instead of optimality as the performance metric may lead to problems with tractable solution. The results of Chapter III and Chapter IV show that signaling strategies based on the common beliefs' supports can efficiently control the systems under consideration. These results suggest that we may be able to stabilize a general class of decentralized systems with bounded uncertainties by proper design of singling based on the supports of common beliefs.

Dynamic networked systems with strategic behavior can only operate according to equilibrium strategies. Proving the existence of appropriate equilibrium strategies and computing these strategies are challenging problems. The results of Chapter V provide an algorithm to compute CIB-PBE for a general class of dynamic games. However, the computational complexity of the algorithm is still high due to the continuous space of CIB beliefs. Thus, we need to develop computational methods to resolve the complexity issue. Sampling or simulation based approaches are potential candidates to approximate CIB beliefs and to compute CIB-PBE. In addition to computing one equilibrium, it is more desirable to compute a Pareto efficient equilibrium. Generally, CIB-PBE is a proper subclass of PBE, and a CIB-PBE may not be Pareto efficient. This is due to the fact that an efficient equilibrium may require informa-

tion in addition to CIB beliefs to coordinate the DMs. In order to achieve Pareto efficiency, we may need to consider information states that include the CIB beliefs and additional information that could be used for efficient coordination. We should be able to compute Pareto efficient equilibria if the additional information needed for efficient coordination has a fixed domain. Identifying such information states is a difficult open problem.

We hope the above thoughts can provide useful starting points for future research that could significantly extend the solution methods and results presented in this thesis.

APPENDICES

APPENDIX A

Appendix for Multi-State Channel Sensing

Proof of Property II.3 .

$$xP - yP = \sum_{i=2}^{K} \left[\left(\sum_{j=i}^{K} (x(j) - y(j)) \right) (P_i - P_{i-1}) \right]. \tag{A.1}$$

Note that $\sum_{j=i}^{K}(x(j)-y(j)) \geq 0$ since $x \geq_{st} y$. Then, by assumption (A1) $P_i \geq_{st} P_{i-1}$ we get

$$\left(\sum_{j=i}^{K} (x(j) - y(j)) \right) (P_i - P_{i-1}) \geq_{st} \mathbf{0}. \tag{A.2}$$

□

Proof of Property II.4 . From Property II.3, (A1) and (A3) we obtain

$$P_i P \leq_{st} P_K P \leq_{st} P_L, \tag{A.3}$$

$$P_i P \geq_{st} P_1 P \geq_{st} P_{L-1}. \tag{A.4}$$

Therefore, (A.3) and (A.4) give

$$P_{L-1} \leq_{st} \sum_{i=1}^{K} x(i)P_iP = xP^2 \leq_{st} P_L. \tag{A.5}$$

□

Proof of Property II.5 . We prove this Property by induction. The Property is true at $t = 0$ by (A2).

Assume the Property is true at t. If n, m are not selected at t, $\pi_{t+1}^n = \pi_t^n P$, $\pi_{t+1}^m = \pi_t^m P$.

By the induction hypothesis we have $\pi_t^n \leq_{st} \pi_t^m$ or $\pi_t^m \leq_{st} \pi_t^n$. Then from Property II.3 we obtain $\pi_{t+1}^n \leq_{st} \pi_{t+1}^m$ or $\pi_{t+1}^m \leq_{st} \pi_{t+1}^n$.

Suppose, without loss of generality, that channel n is selected at t. Since channel m is not selected at t, $\pi_{t+1}^m = \pi_t^m P \in \Delta(S)P^2$. Then from Property II.4 we have either $\pi_{t+1}^n \leq_{st} \pi_{t+1}^m$ or $\pi_{t+1}^m \leq_{st} \pi_{t+1}^n$. □

Proof of Property II.6. (i) Since $x \geq_{st} y$ and $v_i \geq v_{i-1}$, $i = 2, 3, \ldots, K-1$, by summation by parts we have

$$\begin{aligned}
&(x-y)v \\
=&\sum_{i=2}^{K} \left[\left(\sum_{j=i}^{K} (x(j) - y(j)) \right) (v_i - v_{i-1}) \right] \geq 0.
\end{aligned} \tag{A.6}$$

(ii) From the definition of U we have:

$$\text{For } i < L, U_i - U_{i-1} = R_i - R_{i-1}. \tag{A.7}$$

$$\begin{aligned}
\text{For } i \geq L, U_i - U_{i-1} =& R_i - R_{i-1} + \beta(P_i - P_{i-1})U \\
\geq& R_i - R_{i-1}.
\end{aligned} \tag{A.8}$$

Then, for all i, from the definition of M we obtain

$$M_i - M_{i-1} \geq U_i - U_{i-1} \geq R_i - R_{i-1} \geq 0. \tag{A.9}$$

Since $x \geq_{st} y$, from (A.9) and the result of part (i) we have

$$(x-y)M \geq (x-y)U \geq (x-y)R \geq 0. \tag{A.10}$$

(iii) Because of Assumption (A4) and the result of part (ii) we have:

$$\begin{aligned} \text{For } i < L, U_i - U_{i-1} =& R_i - R_{i-1} \\ \geq& \beta(P_i - P_{i-1})M \\ \geq& \beta(P_i - P_{i-1})U. \end{aligned} \tag{A.11}$$

$$\begin{aligned} \text{For } i \geq L, U_i - U_{i-1} =& R_i - R_{i-1} + \beta(P_i - P_{i-1})U \\ \geq& \beta(P_i - P_{i-1})U. \end{aligned} \tag{A.12}$$

Then, (A.11) and (A.12) imply that $U - \beta PU$ is in increasing order, consequently,

$$(x-y)U \geq \beta(x-y)PU. \tag{A.13}$$

Since $M = U + \beta \sum_{i \geq L} p_{Ki} PU$,

$$\begin{aligned} (x-y)M \geq& \beta(x-y)PU + \beta \sum_{i \geq L} p_{Ki} \beta(xP - yP)PU \\ =& \beta(x-y)PM. \end{aligned} \tag{A.14}$$

(iv) If $x(i) = y(i)$ for all $i \geq L$, then $x(i) - y(i) = 0$ for $i \geq L$.

Define $v := (v_1, v_2, \ldots, v_K)$ such that

$$v_i = R_i - \beta P_i M, \quad \text{for } i = 1, 2, \ldots, L-1, \tag{A.15}$$

$$v_i = v_{L-1}, \quad \text{for } i \geq L. \tag{A.16}$$

From assumption (2.26) in (A4) we know that $v_i - v_{i-1} = R_i - R_{i-1} - \beta(P_i - P_{i-1})M \geq 0$ for $i \leq L-1$ and $v_i - v_{i-1} = 0$ for $i \geq L$. Then from the result of part (i) we obtain

$$(x-y)(R - \beta PM) = (x-y)v \geq 0. \tag{A.17}$$

The case where $x(i) = y(i)$ for all $i < L$ can be proved in the same way.

□

Proof of Property II.7. We want to show that under $g_{0:T}^{O_0}$, at any time t the ordering O_t has the property that
$\pi_t^{O_t(1)} \leq_{st} \pi_t^{O_t(2)} \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N)}$.
At $t = 0$, by the statement of Property II.7, the initial ordering O_0 is such that $\pi_0^{O_0(1)} \leq_{st} \pi_0^{O_0(2)} \leq_{st} \cdots \leq_{st} \pi_0^{O_0(N)}$.
Suppose at time t, the ordering O_t is such that $\pi_t^{O_t(1)} \leq_{st} \pi_t^{O_t(2)} \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N)}$.
If the observation is $Y_t \geq L$, the new ordering is $O_{t+1} = \hat{m}(O_t, Y_t) = O_t$ and the PMFs of the channels evolves to

$$\pi_{t+1}^n = \pi_t^n P, \quad \text{for } n \neq O_t(N), \tag{A.18}$$

$$\pi_{t+1}^{O_t(N)} = P_{Y_t} \geq_{st} P_L. \tag{A.19}$$

From Properties II.3 and II.4 we know that

$$\begin{aligned}\pi_t^{O_t(1)}P &\leq_{st} \pi_t^{O_t(2)}P \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N-1)}P \\ &\leq_{st} P_L \leq_{st} P_{Y_t}. \end{aligned} \tag{A.20}$$

On the other hand, if the observation is $Y_t < L$, the new ordering is $O_{t+1} = \hat{m}(O_t, Y_t) = SO_t$ and the PMFs of the channels become

$$\pi_{t+1}^n = \pi_t^n P, \quad \text{for } n \neq O_t(N), \tag{A.21}$$

$$\pi_{t+1}^{O_t(N)} = P_{Y_t} \leq_{st} P_{L-1}. \tag{A.22}$$

Again, from Properties II.3 and II.4 we get

$$P_{Y_t} \leq_{st} P_{L-1} \leq_{st} \pi_t^{O_t(1)}P \leq_{st} \pi_t^{O_t(2)}P \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N-1)}P. \tag{A.23}$$

Thus, the ordering-based policy $g_{0:T}^{O_0}$ selects at any time t the channel $O_t(N)$ from the ordering O_t with $\pi_t^{O_t(1)} \leq_{st} \pi_t^{O_t(2)} \leq_{st} \cdots \leq_{st} \pi_t^{O_t(N)}$. This ordering-based policy is exactly the same as the myopic policy g^m. □

We first establish a lemma that is needed for the proof of Properties II.8-II.11.

Lemma A.1. *The functions $V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$, $t = 1, 2, \ldots, T$ (defined by eq. (2.56)), are linear in every component $\pi_t^n, n = 1, 2, \ldots, N$.*
That is, for all $n = 1, 2, \ldots, N$

$$V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) = \sum_{i=1}^{K} \pi^n(i) V_t(O_t, \pi_t^1, \ldots, \pi_t^{n-1}, e_i, \pi_t^{n+1}, \ldots, \pi_t^N), \tag{A.24}$$

where e_i is the vector with 1 in the ith position and 0 otherwise, i.e.
$e_i = [0, \ldots, 0, \underset{\uparrow \, ith \ position}{1}, 0, \ldots, 0]$.

Furthermore, $L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N)$ *satisfies for* $n = 2, 3, \ldots, N$

$$L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) = \sum_{i=1}^{K} \pi^n(i) L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \ldots, \pi_t^{n-1}, e_i, \pi_t^{n+1}, \ldots, \pi_t^N), \tag{A.25}$$

$$L_t(O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) = \sum_{i=1}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) V_t(O_t, e_i, \pi_t^2, \ldots, \pi_t^N). \tag{A.26}$$

Proof. From the definition of V_t (eq. (2.56)) we have

$$\begin{aligned}
&V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\
&= \sum_{i=1}^{K} \pi_t^n(i) \mathbb{E}^{g_{t:T}^{O_t}} \left[\sum_{s=t}^{T} \beta^{s-t} R(s) | \pi_t^1, \pi_t^2, \ldots, \pi_t^N, X_t^n = i \right] \\
&= \sum_{i=1}^{K} \pi_t^n(i) \mathbb{E}^{g_{t:T}^{O_t}} \left[\sum_{s=t}^{T} \beta^{s-t} R(s) | \pi_t^1, \ldots, \pi_t^{n-1}, \pi_t^{n+1}, \ldots, \pi_t^N, \pi_t^n = e_i \right] \\
&= \sum_{i=1}^{K} \pi_t^n(i) V_t(O_t, \pi_t^1, \ldots, \pi_t^{n-1}, e_i, \pi_t^{n+1}, \ldots, \pi_t^N).
\end{aligned} \tag{A.27}$$

The third equality in (A.27) follows from the specification of the ordering-based policy $g_{t:T}^{O_t}$ and the fact that conditional on $\{X_t^n = i, \pi_t^n\}$ the evolution of channel n is the same as that conditional on $\{\pi_t^n = e_i\}$.

Furthermore, L_t is the difference of two V_t's, so the linearity of V_t leads directly to equations (A.25) and (A.26). □

We proceed now with the proof of Properties II.8-II.11. In the following proof, we use the notation

$$\pi_t^{k_1:k_2} := (\pi_t^{k_1}, \pi_t^{k_1+1}, \ldots, \pi_t^{k_2}), \tag{A.28}$$

$$\pi_t^{k_1:k_2} P := (\pi_t^{k_1} P, \pi_t^{k_1+1} P, \ldots, \pi_t^{k_2} P). \tag{A.29}$$

Proof of Properties II.8-II.11. First note that Property II.10 is a special case of Prop-

erty II.9. This can be seen as follows.

Without loss of generality, let $O_t(n) = 1, O_t(m) = 2$, and $\pi_t^1 \geq_{st} \pi_t^2$. Note that

$$V_t(O_t, \pi_t^2, \pi_t^2, \ldots, \pi_t^N) = V_t(W_{nm}O_t, \pi_t^2, \pi_t^2, \ldots, \pi_t^N). \tag{A.30}$$

Applying Property II.9 at time t, we have

$$\begin{aligned}
&V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) - V_t(W_{nm}O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\
=&V_t(O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) - V_t(O_t, \pi_t^2, \pi_t^2, \ldots, \pi_t^N) \\
&+ V_t(W_{nm}O_t, \pi_t^2, \pi_t^2, \ldots, \pi_t^N) - V_t(W_{nm}O_t, \pi_t^1, \pi_t^2, \ldots, \pi_t^N) \\
=&L_t(O_t, \pi_t^1, \pi_t^2, \pi_t^2, \ldots, \pi_t^N) - L_t(W_{nm}O_t, \pi_t^1, \pi_t^2, \pi_t^2, \ldots, \pi_t^N) \geq 0.
\end{aligned} \tag{A.31}$$

Therefore, Property II.10 is true at time t once Property II.9 is true at time t.

We prove all three Properties II.8, II.9 and II.11 simultaneously by induction. For both the basis of induction and the induction we consider two cases.

(i) When channel 1 is not the right-most channel in O_t (i.e. $n \neq N$ and $O_t(N) \neq 1$).

(ii) When channel 1 is the right-most channel in O_t (i.e. $n = N$ and $O_t(N) = 1$).

Basis of induction

It can be verified that Properties II.8, II.9 and II.11 are true at time $t = T$.

Induction hypothesis

Assume that the assertions of Properties II.8, II.9 and II.11 are true for time $t+1, t+2, \ldots, T$.

Induction step

We prove here Properties II.8, II.9 and II.11 for t.

We first develop five expressions (A.36),(A.38),(A.39), (A.40) and (A.43) for L_t and L_{t+1}, defined by eq. (2.60), that will be useful in the sequel.

For any PMF $\pi \in \Delta(S)$ we define

$$\underline{\pi} := (\pi(1), \pi(2), \ldots, \pi(L-2), \sum_{i=L-1}^{K} \pi(i), 0, \ldots, 0), \tag{A.32}$$

$$\bar{\pi} := (0, \ldots, 0, \sum_{i=1}^{L} \pi(i), \pi(L+1), \ldots, \pi(K)). \tag{A.33}$$

Then, $\underline{\pi}, \bar{\pi} \in \Delta(S)$, and

$$\pi = \underline{\pi} + \bar{\pi} - e_L + \sum_{i=L}^{K} \pi(i)(e_L - e_{L-1}). \tag{A.34}$$

Furthermore, if $\hat{\pi} \geq_{st} \pi$, it follows that

$$\underline{\hat{\pi}} \geq_{st} \underline{\pi}, \quad \bar{\hat{\pi}} \geq_{st} \bar{\pi}. \tag{A.35}$$

Consider any arbitrary ordering $O \in \mathcal{O}$. When $O(N) \neq 1$, assume $O(N) = 2$ without any loss of generality. Then,

$$\begin{aligned}
&L_t(O, \hat{\pi}_t^1, \pi_t^1, \pi_t^{2:N}) \\
=&(\pi_t^2 R - \pi_t^2 R) + \beta \sum_{i<L} \pi_t^2(i)(V_{t+1}(SO, \hat{\pi}_t^1 P, P_i, \pi_t^{3:N} P) - V_{t+1}(SO, \pi_t^1 P, P_i, \pi_t^{3:N} P)) \\
&\quad + \beta \sum_{i \geq L} \pi_t^2(i)(V_{t+1}(O, \hat{\pi}_t^1 P, P_i, \pi_t^{3:N} P) - V_{t+1}(O, \pi_t^1 P, P_i, \pi_t^{3:N} P)) \\
=&\beta \sum_{i<L} \pi_t^2(i) L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) + \beta \sum_{i \geq L} \pi_t^2(i) L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P).
\end{aligned} \tag{A.36}$$

Furthermore, by the induction hypothesis for Property II.8, we get, for all $i = 1, 2, \ldots, K$,

$$L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \geq L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P). \tag{A.37}$$

Therefore,

$$\begin{aligned}\beta L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^{1:N} P) =& \beta \sum_{i=1}^{L} \pi_t^2(i) L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \\ \geq & L_t(O, \hat{\pi}_t^1, \pi_t^{1:N}). \qquad \text{(A.38)} \\ L_t(O, \hat{\pi}_t^1, \pi_t^{1:N}) \geq & \beta \sum_{i=1}^{L} \pi_t^2(i) L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \\ = & \beta L_{t+1}(O, \hat{\pi}_t^1 P, \pi_t^{1:N} P). \qquad \text{(A.39)}\end{aligned}$$

When $O(N) = 1$,

$$\begin{aligned}& L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ :=& V_t(O_t, \hat{\pi}_t^1, \pi_t^{2:N}) - V_t(O_t, \pi_t^1, \pi_t^{2:N}) \\ =& (\hat{\pi}_t^1 R - \pi_t^1 R) + \beta \sum_{i<L} (\hat{\pi}_t^1(i) - \pi_t^1(i)) V_{t+1}(SO_t, P_i, \pi_t^{2:N} P) \\ & + \beta \sum_{i \geq L} (\hat{\pi}_t^1(i) - \pi_t^1(i)) V_{t+1}(O_t, P_i, \pi_t^{2:N} P) \\ =& (\hat{\pi}_t^1 - \pi_t^1) R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) + \beta L_{t+1}(O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\ & + \beta \left[V_{t+1}(O_t, P_L, \pi_t^{2:N} P) - V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) \right] \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) \right]. \qquad \text{(A.40)}\end{aligned}$$

The last equality in (A.40) follows from the linearity of L_t (Lemma A.1) and the definition of $\underline{\pi}, \bar{\pi}$ given by (A.32)-(A.33).

Furthermore, using (A.40) we get

$$
\begin{aligned}
&L_t(O, \hat{\pi}_t^1, \pi_t^{1:N}) - \beta L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
=&(\hat{\pi}_t^1 - \pi_t^1)R + \beta L_{t+1}(SO, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) + \beta L_{t+1}(O, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta \left[V_{t+1}(O, P_L, \pi_t^{2:N} P) - V_{t+1}(SO, P_{L-1}, \pi_t^{2:N} P) \right] \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) \right] \\
&- \beta L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, \pi_t^{2:N} P) \\
=&(\hat{\pi}_t^1 - \pi_t^1)R + \beta L_{t+1}(O, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) - \beta L_{t+1}(SO, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta \left[V_{t+1}(O, P_L, \pi_t^{2:N} P) - V_{t+1}(SO, P_L, \pi_t^{2:N} P) \right] \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) \right] \\
\leq&(\hat{\pi}_t^1 - \pi_t^1)R + \beta(\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)PU \\
&+ \beta \left[V_{t+1}(O, P_L, \pi_t^{2:N} P) - V_{t+1}(SO, P_L, \pi_t^{2:N} P) \right] \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) \right]. \qquad \text{(A.41)}
\end{aligned}
$$

The second equality in (A.41) follows from (A.34) and the linearity of L_t (Lemma A.1). The inequality in (A.41) follows from the induction hypothesis for the upper bound of Property II.8 at $t+1$ and the fact that $\bar{\hat{\pi}}_t^1 P \geq_{st} \bar{\pi}_t^1 P$.

For the last term in (A.41), note that

$$
\begin{aligned}
&V_{t+1}(O, P_L, \pi_t^{2:N} P) - V_{t+1}(SO, P_L, \pi_t^{2:N} P) \\
=&L_{t+1}(O, P_L, P_{L-1}, \pi_t^{2:N} P) - L_{t+1}(SO, P_L, P_{L-1}, \pi_t^{2:N} P) + V_{t+1}(O, P_{L-1}, \pi_t^{2:N} P) \\
&- V_{t+1}(W_{12} \cdots W_{(N-1)(N-2)} W_{N(N-1)} O, P_{L-1}, \pi_t^{2:N} P) \\
\leq&L_{t+1}(O, P_L, P_{L-1}, \pi_t^{2:N} P) - L_{t+1}(SO, P_L, P_{L-1}, \pi_t^{2:N} P) \\
\leq&(P_L - P_{L-1})U. \qquad \text{(A.42)}
\end{aligned}
$$

The equality in (A.42) follows from the definition of L_{t+1} and the fact that $SO = W_{12} \cdots W_{(N-1)(N-2)} W_{N(N-1)} O$. The inequalities in (A.42) follow by the induction hypothesis for Property II.10 and Property II.8 at $t+1$.

Therefore, using (A.42) and (A.41) we get

$$\begin{aligned}
&L_t(O, \hat{\pi}_t^1, \pi_t^1, \pi_t^{2:N}) - \beta L_{t+1}(SO, \hat{\pi}_t^1 P, \pi_t^1 P, \pi_t^{2:N} P) \\
\leq &(\hat{\pi}_t^1 - \pi_t^1)R + \beta(\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)PU + \beta(P_L - P_{L-1})U \sum_{i=L}^{K}(\hat{\pi}_t^1(i) - \pi_t^1(i)) \\
= &(\hat{\pi}_t^1 - \pi_t^1)U. \qquad \text{(A.43)}
\end{aligned}$$

The last equality in (A.43) follows from the definition of the vector U.

Induction step for Property II.8:

We first consider the lower bound of Property II.8.

(i) When $O_t(N) \neq 1$ (i.e. $n \neq N$), we also have $S^{-m}O_t(N) = O_t(m) \neq 1$. Then,

$$\begin{aligned}
L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \geq &\beta L_{t+1}(O_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
\geq &\beta L_{t+1}(S^{1-m}O_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
\geq &L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}). \qquad \text{(A.44)}
\end{aligned}$$

The first inequality in (A.44) follows from (A.38) and the fact that $O_t(N) \neq 1$. The second inequality in (A.44) follows from the induction hypothesis for Property II.8 at $t+1$. The last inequality in (A.44) follows from (A.39) and the fact that $S^{-m}O_t(N) \neq 1$.

(ii) When $O_t(N) = 1$ (i.e. $n = N$).

Since $S^{-m}O_t(N) = O_t(m) \neq 1$, from (A.39) we get

$$\begin{aligned}
&L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
\leq &\beta L_{t+1}(S^{1-m}O_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
= &\beta L_{t+1}(S^{1-m}O_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) + \beta L_{t+1}(S^{1-m}O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta \sum_{i=L}^{K}(\hat{\pi}_t^1(i) - \pi_t^1(i)) L_{t+1}(S^{1-m}O_t, P_L, P_{L-1}, \pi_t^{2:N} P). \qquad \text{(A.45)}
\end{aligned}$$

Since $O_t(N) = 1$, applying (A.40) we obtain

$$
\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
=&(\hat{\pi}_t^1 - \pi_t^1)R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) + \beta L_{t+1}(O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta \left[V_{t+1}(O_t, P_L, \pi_t^{2:N} P) - V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) \right] \sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) \\
&- L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^1, \pi_t^{2:N}) \\
\geq&(\hat{\pi}_t^1 - \pi_t^1)R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) - \beta L_{t+1}(S^{1-m}O_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta L_{t+1}(O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) - \beta L_{t+1}(S^{1-m}O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta \left[V_{t+1}(O_t, P_L, \pi_t^{2:N} P) - V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) \right. \\
&\quad \left. - L_{t+1}(S^{1-m}O_t, P_L, P_{L-1}, \pi_t^{2:N} P) \right] \sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)) \\
\geq&(\hat{\pi}_t^1 - \pi_t^1)R + \beta L_{t+1}(SO_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) - \beta L_{t+1}(S^{1-m}O_t, \underline{\hat{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\
&+ \beta \left[V_{t+1}(O_t, P_L, \pi_t^{2:N} P) - V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) \right. \\
&\quad \left. - L_{t+1}(S^{1-m}O_t, P_L, P_{L-1}, \pi_t^{2:N} P) \right] \sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i)). \qquad \text{(A.46)}
\end{aligned}
$$

The equality in (A.46) follows from (A.40) and the fact that $O_t(N) = 1$. The first inequality in (A.46) follows from (A.45). The second inequality in (A.46) follows from the induction hypothesis for Property II.8 at $t + 1$.

Letting $\underline{O}_{t+1} := S^{1-m}O_t$ and $\underline{n} := N + 1 - m, \underline{m} := N - m$, we have $\underline{m} < \underline{n}$ and

$$
\underline{O}_{t+1}(\underline{n}) = S^{1-m}O_t(\underline{n}) = 1, \quad SO_t = S^{-(\underline{m})}\underline{O}_{t+1}. \qquad \text{(A.47)}
$$

Consequently, the induction hypothesis for the upper bound of Property II.8 at $t + 1$

gives

$$
\begin{aligned}
&L_{t+1}(S^{1-m}O_t, \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) - L_{t+1}(SO_t, \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\
=&L_{t+1}(\underline{O}_{t+1}, \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) - L_{t+1}(S^{-(\underline{m})}\underline{O}_{t+1}, \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\
\leq&(\hat{\underline{\pi}}_t^1 P - \underline{\pi}_t^1 P)U. \qquad \text{(A.48)}
\end{aligned}
$$

Letting $\underline{m}' := 1$, we have $\underline{m}' < n = N$ and $A_{\underline{m}'n}O_t = SO_t$. Therefore,

$$
\begin{aligned}
&V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) - V_{t+1}(O_t, P_L, \pi_t^{2:N} P) + L_{t+1}(S^{1-m}O_t, P_L, P_{L-1}, \pi_t^{2:N} P) \\
\leq&V_{t+1}(SO_t, P_{L-1}, \pi_t^{2:N} P) - V_{t+1}(O_t, P_L, \pi_t^{2:N} P) + L_{t+1}(O_t, P_L, P_{L-1}, \pi_t^{2:N} P) \\
=&V_{t+1}(A_{\underline{m}'n}O_t, P_{L-1}, \pi_t^{2:N} P) - V_{t+1}(O_t, P_{L-1}, \pi_t^{2:N} P) \\
\leq&h - P_{L-1}R. \qquad \text{(A.49)}
\end{aligned}
$$

The first inequality in (A.49) follows from the induction hypothesis for the lower bound of Property II.8 at $t+1$. The last inequality in (A.49) follows from the induction hypothesis for Property II.11 at $t+1$.

Using (A.48) and (A.49) in (A.46) we obtain

$$
\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
\geq&(\hat{\pi}_t^1 - \pi_t^1)R - \beta(\hat{\underline{\pi}}_t^1 P - \underline{\pi}_t^1 P)U - \beta \sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i))(h - P_{L-1}R) \\
=&(\hat{\underline{\pi}}_t^1 - \underline{\pi}_t^1)(R - \beta U) + (\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R + \sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i))(R_L - R_{L-1} - \beta(h - P_{L-1}R)) \\
\geq&0. \qquad \text{(A.50)}
\end{aligned}
$$

The last inequality in (A.50) is true because: the terms $(\hat{\underline{\pi}}_t^1 - \underline{\pi}_t^1)(R - \beta U)$ and $(\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R$ are positive by parts (iv) and (ii) of Property II.6; the term $(R_L - R_{L-1} - \beta(h - P_{L-1}R))$ is positive by Condition (A4).

The proof of the lower bound of Property II.8 is now complete.

Now consider the upper bound of Property II.8.

Let $O'_t := S^{N-n}O_t;$, then $O'_t(N) = 1$ and $SO'_t(1) = 1$. Consequently,

$$\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(S^{-m}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
\leq &L_t(O'_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(SO'_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
\leq &L_t(O'_t, \hat{\pi}_t^1, \pi_t^{1:N}) - \beta L_{t+1}(SO'_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
\leq &(\hat{\pi}_t^1 - \pi_t^1)U. \qquad \text{(A.51)}
\end{aligned}$$

The first inequality in (A.51) is true because of the lower bound of Property II.8 at t. The second inequality in (A.51) follows from (A.39) and the fact that $SO'_t(N) \neq 1$. The third inequality in (A.51) follows from (A.43) and the fact that $O'_t(N) = 1$.

This completes the proof of Property II.8 at time t.

Induction step for Property II.9:

(i) When $O_t(N) \neq 1$ (i.e. $n \neq N$), assume $O_t(N) = 2$ without loss of generality. Then because of (A.36),

$$\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
=&\beta \sum_{i<L} \pi^2(i) \left[L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \right. \\
&\left. - L_{t+1}(W_{(n+1)(m+1)}(SO_t), \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \right] \\
&+ \beta \sum_{i \geq L} \pi^2(i) \left[L_{t+1}(O_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \right. \\
&\left. - L_{t+1}(W_{nm}O_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \right]. \qquad \text{(A.52)}
\end{aligned}$$

By the induction hypothesis for Property II.9, each term in (A.52) is positive and

smaller than $(\hat{\pi}_t^1 P - \pi_t^1 P)M$. Thus,

$$\begin{aligned} 0 \leq & L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ \leq & \beta(\hat{\pi}_t^1 P - \pi_t^1 P)M \leq (\hat{\pi}_t^1 - \pi_t^1)M. \end{aligned} \tag{A.53}$$

The last inequality in (A.53) holds by part (iii) of Property II.6.

(ii) $O_t(N) = 1$ (i.e. $n = N$).

We first consider the lower-bound. Using (A.34) and the linearity of L_t (Lemma A.1) we get

$$\begin{aligned} & L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\ = & L_t(O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) \\ & + L_t(O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N}) \\ & + \left[L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N})\right] \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i))\right]. \end{aligned} \tag{A.54}$$

We consider each of the terms

(a) $L_t(O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \underline{\hat{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N})$.

(b) $L_t(O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \hat{\pi}_t^1, \pi_t^{2:N})$.

(c) $\left[L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N})\right] \left[\sum_{i=L}^{K} (\hat{\pi}_t^1(i) - \pi_t^1(i))\right]$.

that appear in the right hand side of (A.54) separately.

(a) Consider the first term.

Let $O_t' = S(W_{Nm}O_t) = W_{1m+1}(SO_t)$, then $O_t'(m+1) = 1$ and $W_{m+1,1}O_t' = SO_t$.

Therefore,

$$
\begin{aligned}
&L_t(O_t, \hat{\underline{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \hat{\underline{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) \\
=&(\hat{\underline{\pi}}_t^1 - \underline{\pi}_t^1)R + \beta L_{t+1}(SO_t, \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) - L_t(W_{Nm}O_t, \hat{\underline{\pi}}_t^1, \underline{\pi}_t^1, \pi_t^{2:N}) \\
\geq&(\hat{\underline{\pi}}_t^1 - \underline{\pi}_t^1)R + \beta L_{t+1}(SO_t, \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) - \beta L_{t+1}(S(W_{Nm}O_t), \hat{\underline{\pi}}_t^1 P, \underline{\pi}_t^1 P, \pi_t^{2:N} P) \\
\geq&(\hat{\underline{\pi}}_t^1 - \underline{\pi}_t^1)R - \beta(\hat{\underline{\pi}}_t^1 P - \underline{\pi}_t^1 P)M \geq 0. \qquad \text{(A.55)}
\end{aligned}
$$

The first equality in (A.55) follows from (A.40) and that fact that $\hat{\underline{\pi}}_t^1(i) = \underline{\pi}_t^1(i) = 0$ for $i \geq L$. The first inequality in (A.55) follows from (A.38). The second inequality in (A.55) follows from the induction hypothesis for the upper bound of Property II.9 at $t+1$. The last inequality in (A.55) holds by part (iv) of Property II.6.

(b) Consider the second term. From the induction hypothesis for Property II.8 at $t+1$ and the fact that $SO_t = S^{-(N-1)}O_t$ and $O_t(N) = 1$, we obtain.

$$
\begin{aligned}
&L_t(O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}) \\
=&(\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R + \beta L_{t+1}(O_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}) \\
\geq&(\bar{\hat{\pi}}_t^1 - \bar{\pi}_t^1)R + \beta L_{t+1}(SO_t, \bar{\hat{\pi}}_t^1 P, \bar{\pi}_t^1 P, \pi_t^{2:N} P) - L_t(W_{Nm}O_t, \bar{\hat{\pi}}_t^1, \bar{\pi}_t^1, \pi_t^{2:N}). \qquad \text{(A.56)}
\end{aligned}
$$

Then, similar to the first term (a), the second term is positive.

(c) Consider the third term.

Assume $O_t(m) = 2$ without any loss of generality. Then $W_{Nm}O_t(N) = 2$. Therefore,

$$\begin{aligned}
&L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N}) \\
=&R_L - R_{L-1} + \beta \sum_{i<L} \pi^2(i) \left[V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N}P) \right. \\
&\left. - L_{t+1}(SW_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N}P) \right] \\
&+ \beta \sum_{i \geq L} \pi^2(i) \left[V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N}P) \right. \\
&\left. - L_{t+1}(W_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N}P) \right]. \qquad \text{(A.57)}
\end{aligned}$$

Let $O'_t := S(W_{Nm}O_t) = W_{1m+1}(SO_t)$; then $O'_t(m+1) = 1$ and $W_{m+1,1}O'_t = SO_t$. For each term in the first sum in (A.57), we have $P_{L-1} \geq_{st} P_i$ ($i < L$ in the first sum in (A.57)). Therefore, by the induction hypothesis for Property II.10 at $t+1$ we get

$$\begin{aligned}
&V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N}P) \\
&- L_{t+1}(SW_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N}P) \\
\geq& V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(O'_t, P_L, P_i, \pi_t^{3:N}P). \qquad \text{(A.58)}
\end{aligned}$$

Furthermore, since $P_L \geq_{st} \pi_t^{O_t(l)}P$ for all $l = 1, 2, \ldots, N$ by Property II.4, repeatedly applying Property II.10 at $t+1$ we obtain

$$V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) \geq V_{t+1}(W_{(m+2)(m+1)} \cdots W_{N(N-1)}O_t, P_L, P_i, \pi_t^{3:N}P). \qquad \text{(A.59)}$$

Note that $W_{(m+2)(m+1)} \cdots W_{N(N-1)}O_t = A_{N(m+1)}O_t$ and $A_{m1}(A_{N(m+1)}O_t) = S(W_{Nm}O_t) =$

O'_t. Consequently, the induction hypothesis for Property II.11 at $t+1$ gives

$$\begin{aligned}
&V_{t+1}(W_{(m+2)(m+1)}\cdots W_{N(N-1)}O_t, P_L, P_i, \pi_t^{3:N}P)\\
&- V_{t+1}(O'_t, P_L, P_i, \pi_t^{3:N}P)\\
=&V_{t+1}(A_{N(m+)1}O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(A_{m1}(A_{N(m+1)}O_t), P_L, P_i, \pi_t^{3:N}P)\\
\geq& -(h - P_iP^{N-m}R). \qquad \text{(A.60)}
\end{aligned}$$

For each term in the second sum in (A.57), we have

$$\begin{aligned}
&V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N}P)\\
&- L_{t+1}(W_{Nm}O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N}P)\\
\geq&V_{t+1}(O_t, P_L, P_i, \pi_t^{3:N}P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N}P)\\
&- L_{t+1}(O_t, P_L, P_{L-1}, P_i, \pi_t^{3:N}P)\\
=&V_{t+1}(O_t, P_{L-1}, P_i, \pi_t^{3:N}P) - V_{t+1}(SO_t, P_{L-1}, P_i, \pi_t^{3:N}P)\\
\geq& -(h - P_{L-1}R). \qquad \text{(A.61)}
\end{aligned}$$

The inequalities in (A.61) follows from the induction hypothesis at $t+1$ for the lower bound of Property II.9 and Property II.11 respectively.

Using the lower bounds provided by (A.60) and (A.61) for terms in (A.57), we obtain

$$\begin{aligned}
&L_t(O_t, e_L, e_{L-1}, \pi_t^{2:N}) - L_t(W_{Nm}O_t, e_L, e_{L-1}, \pi_t^{2:N})\\
\geq&R_L - R_{L-1} - \beta\sum_{i<L}\pi_t^2(i)(h - P_iP^{N-m}R) - \beta\sum_{i\geq L}\pi_t^2(i)(h - P_{L-1}R)\\
\geq&R_L - R_{L-1} - \beta(h - P_{L-1}R) \geq 0. \qquad \text{(A.62)}
\end{aligned}$$

The second and the last inequalities in (A.62) follows from part (ii) of Property II.6 and condition (A4) respectively.

Since the three terms (a), (b) and (c) in (A.54) are positive, the proof for the lower bound of Property II.9 is complete when $O_t(N) = 1$ (case (ii)).

We now proceed to establish the upper bound of Property II.9 when $O_t(N) = 1$ (case (ii)). Assume $O_t(m) = 2$ without any loss of generality; then $W_{Nm}O_t(N) = 2$. Therefore,

$$
\begin{aligned}
&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
=&L_t(O_t, \hat{\pi}_t^1, \pi_t^{1:N}) - \beta L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) \\
&+ \beta L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
\leq&(\hat{\pi}_t^1 - \pi_t^1)U + \beta L_{t+1}(SO_t, \hat{\pi}_t^1 P, \pi_t^{1:N} P) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
\leq&(\hat{\pi}_t^1 - \pi_t^1)U + \beta L_{t+1}(S(W_{Nm}O_t), \hat{\pi}_t^1 P, \pi_t^{1:N} P) - L_t(W_{Nm}O_t, \hat{\pi}_t^1, \pi_t^{1:N}) \\
=&(\hat{\pi}_t^1 - \pi_t^1)U + \beta \sum_{i \geq L} \pi_t^2(i) \left[L_{t+1}(S(W_{Nm}O_t), \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \right. \\
&\qquad \left. - L_{t+1}(W_{Nm}O_t, \hat{\pi}_t^1 P, \pi_t^1 P, P_i, \pi_t^{3:N} P) \right] \\
\leq&(\hat{\pi}_t^1 - \pi_t^1)U + \beta \sum_{i \geq L} \pi_t^2(i)(\hat{\pi}_t^1 P - \pi_t^1 P)U \\
\leq&(\hat{\pi}_t^1 - \pi_t^1)U + \beta \sum_{i \geq L} p_{Ki}(\hat{\pi}_t^1 P - \pi_t^1 P)U \\
=&(\hat{\pi}_t^1 - \pi_t^1)M. \qquad \text{(A.63)}
\end{aligned}
$$

The first inequality in (A.63) follows from (A.43). The second inequality in (A.63) follows from the induction hypothesis for the lower bound of Property II.9 at $t + 1$. The second equality in (A.63) follows from (A.36). The third inequality in (A.63) follows from the induction hypothesis for Property II.8 and the fact that $\hat{\pi}_t^1 P \geq_{st} \pi_t^1 P$ (since $\hat{\pi}_t^1 \geq_{st} \pi_t^1$ and Property II.3). The last inequality in (A.63) is true because $\pi_t^2 \leq_{st} P_K$. The last equality in (A.63) follows from the definition of M.

The proof of the upper bound of Property II.9 at t is now complete. The proof of the induction step for Property II.9 at t is also complete.

Induction step for Property II.11:

(i) When $O_t(N) \neq 1$ (i.e. $n \neq N$), assume $O_t(N) = N$ without loss of generality. Then,

$$\begin{aligned}
&V_t(A_{nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N}) \\
=&\sum_{i<L} \pi_t^N(i) \left[V_{t+1}(S(A_{nm}O_t), \pi_t^{1:N-1}P, P_i) - V_{t+1}(SO_t, \pi_t^{1:N-1}P, P_i)\right] \\
&+ \sum_{i\geq L} \pi_t^N(i) \left[V_{t+1}(A_{nm}O_t, \pi_t^{1:N-1}P, P_i) - V_{t+1}(O_t, \pi_t^{1:N-1}P, P_i)\right] \\
\leq&\sum_{i<L} \pi_t^N(i) \left(h - \pi_t^1 P(P^{N-n-1}R)\right) + \sum_{i\geq L} \pi_t^N(i) \left(h - \pi_t^1 P(P^{N-n}R)\right) \\
\leq&h - \pi_t^1 P^{N-n}R. \qquad \text{(A.64)}
\end{aligned}$$

The inequalities in (A.64) follows from the induction hypothesis for Property II.11 and part (ii) of Property II.6 respectively.

(ii) When $O_t(N) = 1$ (i.e. $n = N$), assume $O_t(N-1) = N$ without loss of generality. Then $A_{Nm}O_t(N) = O_t(N-1) = N$.

By the recursive equation and the linearity of the function V_{t+1} (eq. (2.58) and Lemma A.1) we obtain

$$\begin{aligned}
&V_t(A_{Nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N}) \\
=&(\pi_t^N - \pi_t^1)R \\
&+ \beta \sum_{i<L} \pi_t^N(i) \left[V_{t+1}(S(A_{Nm}O_t), \pi_t^{1:N-1}P, P_i) - V_{t+1}(A_{Nm}O_t, \pi_t^{1:N-1}P, P_i)\right] \\
&+ \beta \sum_{i<L} \pi_t^1(i) \left[V_{t+1}(A_{Nm}O_t, P_i, \pi_t^{2:N}P) - V_{t+1}(SO_t, P_i, \pi_t^{2:N}P)\right] \\
&+ \beta \sum_{i\geq L} \pi_t^1(i) \left[V_{t+1}(A_{Nm}O_t, P_i, \pi_t^{2:N}P) - V_{t+1}(O_t, P_i, \pi_t^{2:N}P)\right]. \qquad \text{(A.65)}
\end{aligned}$$

Furthermore, each term in the second and the third sum in (A.65) is negative from

repeatedly using Property II.10 at $t+1$. Therefore,

$$\begin{aligned}
&V_t(A_{Nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N}) \\
\leq&(\pi_t^N - \pi_t^1)R \\
&+ \beta \sum_{i<L} \pi_t^N(i) \left[V_{t+1}(S(A_{Nm}O_t), \pi_t^{1:N-1}P, P_i) - V_{t+1}(A_{Nm}O_t, \pi_t^{1:N-1}P, P_i)\right] \\
\leq&(\pi_t^N - \pi_t^1)R + \beta \sum_{i<L} \pi_t^N(i)(h - P_iR) \\
=&\pi_t^N v - \pi_t^1 R. \qquad \text{(A.66)}
\end{aligned}$$

The second inequality in (A.66) follows from the induction hypothesis for Property II.11 and v is the vector such that

$$v_i = \begin{cases} R_i + \beta(h - P_iR), & \text{for } i < L, \\ R_i, & \text{for } i \geq L. \end{cases} \qquad \text{(A.67)}$$

It can be verified that v_i increases with i. Then, from part (i) of Property II.6 and the fact that $\pi_t^N \leq_{st} P_K$ we obtain

$$V_t(A_{Nm}O_t, \pi_t^{1:N}) - V_t(O_t, \pi_t^{1:N}) \ \leq \pi_t^N v - \pi_t^1 R \leq P_K v - \pi_t^1 R = h - \pi_t^1 R. \qquad \text{(A.68)}$$

The last equality in (A.68) follows from the definition of h.

This completes the proof of the induction step for Property II.11 at t, and the proof of the entire induction step. □

APPENDIX B

Appendix for Decentralized Routing

Proof of Lemma III.4. Since there is one possible arrival to any queue and one possible departure from any queue at each time instant, (3.31) holds.

When $(U_t^{1,\hat{g}}, U_t^{2,\hat{g}}) = (0,0)$, both $\overline{X}_t^{1,\hat{g}}$ and $\overline{X}_t^{2,\hat{g}}$ are below the threshold and no customers are routed form any queue. Therefore, the upper bound of the queue lengths at $t+1$ is

$$UB_{t+1}^{\hat{g}} = \lceil TH_t \rceil - 1. \tag{B.1}$$

Moreover, the lower bound of the queue lengths at $t+1$ is the same as the lower bound of $\overline{X}_t^{1,\hat{g}}, \overline{X}_t^{2,\hat{g}}$. That is,

$$LB_{t+1}^{\hat{g}} = \overline{LB}_t^{\hat{g}}. \tag{B.2}$$

When $(U_t^{1,\hat{g}}, U_t^{2,\hat{g}}) = (1,1)$, both $\overline{X}_t^{1,\hat{g}}$ and $\overline{X}_t^{2,\hat{g}}$ are greater than or equal to the threshold. Since the routing only exchanges two customers between the two queues, the queue lengths remain the same as the queue lengths before routing. As a result,

the upper bound and lower bound of the queue lengths at $t+1$ are given by

$$UB_{t+1}^{\hat{g}} = \overline{UB}_t^{\hat{g}}. \tag{B.3}$$

$$LB_{t+1}^{\hat{g}} = \lceil TH_t \rceil. \tag{B.4}$$

When $(U_t^{i,\hat{g}}, U_t^{j,\hat{g}}) = (1,0), i \neq j$, $\overline{X}_t^{i,\hat{g}}$ is greater than or equal to the threshold; $\overline{X}_t^{j,\hat{g}}$ is below the threshold. Since one customer is routed from Q_i to Q_j,

$$X_{t+1}^{i,\hat{g}} = \overline{X}_t^{i,\hat{g}} - 1, \tag{B.5}$$

$$X_{t+1}^{j,\hat{g}} = \overline{X}_t^{j,\hat{g}} + 1. \tag{B.6}$$

Therefore, the upper bound of the queue lengths at $t+1$ becomes

$$\begin{aligned} UB_{t+1}^{\hat{g}} &= \max\left\{\overline{UB}_t^{i,\hat{g}} - 1, \lceil TH_t \rceil - 1 + 1\right\} \\ &= \max\left\{\overline{UB}_t^{i,\hat{g}} - 1, \lceil TH_t \rceil\right\}, \end{aligned} \tag{B.7}$$

and lower bound of the queue lengths at $t+1$ is given by

$$LB_{t+1}^{\hat{g}} = \min\left\{\lceil TH_t \rceil - 1, \overline{LB}_t^{j,\hat{g}} + 1\right\}. \tag{B.8}$$

□

Proof of Lemma III.8. The proof is done by induction.

At time $t = 0$, $X_0^{1,\hat{g}} + X_0^{2,\hat{g}} = X_0^{1,g} + X_0^{2,g} = x_0$.

Suppose the lemma is true at time t.

At time $t+1$, from the system dynamics (3.1)-(3.3) we get, for any g,

$$\begin{aligned} & X_{t+1}^{1,g} + X_{t+1}^{2,g} \\ =& \left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ + A_t^1 + A_t^2. \end{aligned} \tag{B.9}$$

Therefore, it suffices to show that

$$\left(X_t^{1,\hat{g}} - D_t^1\right)^+ + \left(X_t^{2,\hat{g}} - D_t^2\right)^+ \leq_{st} \left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+. \tag{B.10}$$

Consider any realization $(X_t^{1,g}, X_t^{2,g}) = (x^1, x^2)$.

If $x^1, x^2 > 0$, then $\left\lfloor \frac{1}{2}(x^1+x^2) \right\rfloor, \left\lceil \frac{1}{2}(x^1+x^2) \right\rceil > 0$. Therefore,

$$\begin{aligned} & \left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ \\ =& x^1 + x^2 - D_t^1 - D_t^2 \\ =& \left(\left\lfloor \frac{1}{2}(x^1+x^2) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(x^1+x^2) \right\rceil - D_t^2\right)^+. \end{aligned} \tag{B.11}$$

If $x^i = 0$ and $x^j \geq 2$ $(i \neq j)$, then $\left\lfloor \frac{1}{2}(x^1+x^2) \right\rfloor > 0$ and $\left\lceil \frac{1}{2}(x^1+x^2) \right\rceil > 0$. Therefore,

$$\begin{aligned} & \left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ \\ =& x^j - D_t^j \\ \geq& x^1 + x^2 - D_t^1 - D_t^2 \\ =& \left(\left\lfloor \frac{1}{2}(x^1+x^2) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(x^1+x^2) \right\rceil - D_t^2\right)^+. \end{aligned} \tag{B.12}$$

If $x^i = 0$ and $x^j = 1$ $(i \neq j)$, then $\left\lfloor \frac{1}{2}(x^1 + x^2) \right\rfloor = 0$ and $\left\lceil \frac{1}{2}(x^1 + x^2) \right\rceil = 1$. Therefore,

$$\begin{aligned}
&\left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ \\
=&1 - D_t^j \\
\geq_{st}&1 - D_t^2 \\
=&\left(\left\lfloor \frac{1}{2}(x^1 + x^2) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(x^1 + x^2) \right\rceil - D_t^2\right)^+, \qquad \text{(B.13)}
\end{aligned}$$

If $x^1, x^2 = 0$, then $\left\lfloor \frac{1}{2}(x^1 + x^2) \right\rfloor, \left\lceil \frac{1}{2}(x^1 + x^2) \right\rceil = 0$. Therefore,

$$\begin{aligned}
&\left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ \\
=&0 \\
=&\left(\left\lfloor \frac{1}{2}(x^1 + x^2) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(x^1 + x^2) \right\rceil - D_t^2\right)^+. \qquad \text{(B.14)}
\end{aligned}$$

As a result of (B.11)-(B.14), we obtain

$$\begin{aligned}
&\left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ \\
\geq_{st}&\left(\left\lfloor \frac{1}{2}(X_t^{1,g} + X_t^{2,g}) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(X_t^{1,g} + X_t^{2,g}) \right\rceil - D_t^2\right)^+. \qquad \text{(B.15)}
\end{aligned}$$

Then, from (B.15), the induction hypothesis and Corollary III.7 we obtain

$$\begin{aligned}
&\left(X_t^{1,g} - D_t^1\right)^+ + \left(X_t^{2,g} - D_t^2\right)^+ \\
\geq_{st}&\left(\left\lfloor \frac{1}{2}(X_t^{1,g} + X_t^{2,g}) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(X_t^{1,g} + X_t^{2,g}) \right\rceil - D_t^2\right)^+ \\
\geq_{st}&\left(\left\lfloor \frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}}) \right\rfloor - D_t^1\right)^+ + \left(\left\lceil \frac{1}{2}(X_t^{1,\hat{g}} + X_t^{2,\hat{g}}) \right\rceil - D_t^2\right)^+ \\
=&\left(\min(X_t^{1,\hat{g}}, X_t^{2,\hat{g}}) - D_t^1\right)^+ + \left(\max(X_t^{1,\hat{g}}, X_t^{2,\hat{g}}) - D_t^2\right)^+ \\
\geq_{st}&\left(X_t^{1,\hat{g}} - D_t^1\right)^+ + \left(X_t^{2,\hat{g}} - D_t^2\right)^+. \qquad \text{(B.16)}
\end{aligned}$$

The first and second stochastic inequalities in (B.16) follow from (B.15) and the induction hypothesis, respectively. The equality in (B.16) follows from Corollary III.7. The last stochastic inequality in (B.16) is true because D_t^1, D_t^2 are i.i.d. and independent of $X_t^{1,\hat{g}}, X_t^{2,\hat{g}}$.

Thus, inequality (B.10) is true, and the proof of the lemma is complete. □

Proof of Lemma III.15. The proof is done by induction. At $t = 0$, (3.67), (3.68) and (3.69) hold if we let $Y_0^i = X_0^{i,g_0}$ for $i = 1, 2$.

Assume the assertion of this lemma is true at time t; we want to show that the assertion is also true at time $t + 1$.

For that matter we claim the following.

Claim 1

$$X_{t+1}^{1,\hat{g}} + X_{t+1}^{2,\hat{g}} = \overline{X}_t^{1,\hat{g}} + \overline{X}_t^{2,\hat{g}} \quad a.s., \tag{B.17}$$

$$\max_i \left(X_{t+1}^{i,\hat{g}} \right) \leq \max_i \left(\overline{X}_t^{i,\hat{g}} \right) \quad a.s. \tag{B.18}$$

Claim 2

There exists $Y_{t+1}^i, i = 1, 2$ such that

$$\mathbb{P}\left(Y_{t+1}^i = y_{t+1} | Y_{0:t}^i = y_{0:t}\right) = \mathbb{P}\left(X_{t+1}^{i,g0} = y_{t+1} | X_{0:t}^{i,g_0} = y_{0:t}\right) \text{ for all } y_{0:t}, \tag{B.19}$$

$$\overline{X}_t^{1,\hat{g}} + \overline{X}_t^{2,\hat{g}} \leq Y_{t+1}^1 + Y_{t+1}^2 \quad a.s., \tag{B.20}$$

$$\max_i \left(\overline{X}_t^{i,\hat{g}} \right) \leq \max_i \left(Y_{t+1}^i \right) \quad a.s. \tag{B.21}$$

We assume the above claims to be true and prove them after the completion of the proof of the induction step.

For all $y_{0:t+1}$, from (B.19) and the induction hypothesis for (3.67) we get for $i = 1, 2$

$$\begin{aligned}
&\mathbb{P}\left(Y^i_{0:t+1} = y_{0:t+1}\right) \\
=&\mathbb{P}\left(Y^i_{t+1} = y_{t+1} | Y^i_{0:t} = y_{0:t}\right) \mathbb{P}\left(Y^i_t = y_t, \ldots, Y^i_0 = y_0\right) \\
=&\mathbb{P}\left(X^{i,g0}_{t+1} = y_{t+1} | X^{i,g_0}_{0:t} = y_{0:t}\right) \mathbb{P}\left(X^{i,g_0}_{0:t} = y_{0:t}\right) \\
=&\mathbb{P}\left(X^{i,g0}_{0:t+1} = y_{0:t+1}\right). \qquad \text{(B.22)}
\end{aligned}$$

From (B.17) and (B.20) we obtain

$$\begin{aligned}
X^{1,\hat{g}}_{t+1} + X^{2,\hat{g}}_{t+1} =& \overline{X}^{1,\hat{g}}_t + \overline{X}^{2,\hat{g}}_t \\
\leq& Y^1_{t+1} + Y^2_{t+1} \quad a.s. \qquad \text{(B.23)}
\end{aligned}$$

Furthermore, combination of (B.18) and (B.21) gives

$$\max_i \left(X^{i,\hat{g}}_{t+1}\right) \leq \max_i \left(\overline{X}^{i,\hat{g}}_t\right) = \max_i \left(Y^i_{t+1}\right) \quad a.s. \qquad \text{(B.24)}$$

Therefore, the assertions (3.67), (3.68) and (3.69) of the lemma are true at $t+1$ by (B.22), (B.23) and (B.24), respectively.

We now prove claims 1 and 2.

Proof of Claim 1

From the system dynamics (3.1)-(3.2)

$$X^{1,\hat{g}}_{t+1} = \overline{X}^{i,\hat{g}}_t - U^{i,\hat{g}}_t + U^{j,\hat{g}}_t, \qquad \text{(B.25)}$$

$$X^{2,\hat{g}}_{t+1} = \overline{X}^{i,\hat{g}}_t - U^{i,\hat{g}}_t + U^{j,\hat{g}}_t. \qquad \text{(B.26)}$$

Therefore, (B.17) follows by summing (B.25) and (B.26).

For (B.18), consider $X^{1,\hat{g}}_{t+1}$ (the case of $X^{2,\hat{g}}_{t+1}$ follows from similar arguments).

When $U_t^{2,\hat{g}} = 0$,

$$X_{t+1}^{1,\hat{g}} = \overline{X}_t^{1,\hat{g}} - U_t^{1,\hat{g}} \leq \max_i \left(\overline{X}_t^{i,\hat{g}} \right). \tag{B.27}$$

When $U_t^{1,\hat{g}} = U_t^{2,\hat{g}} = 1$,

$$X_{t+1}^{1,\hat{g}} = \overline{X}_t^{1,\hat{g}} \leq \max_i \left(\overline{X}_t^{i,\hat{g}} \right). \tag{B.28}$$

When $U_t^{1,\hat{g}} = 0, U_t^{2,\hat{g}} = 1$, $\overline{X}_t^{1,\hat{g}}$ is less than the threshold and $\overline{X}_t^{2,\hat{g}}$ is greater than or equal to the threshold. Therefore, by (B.25),

$$\begin{aligned} X_{t+1}^{1,\hat{g}} = \overline{X}_t^{1,\hat{g}} + 1 \leq & \lceil TH_t \rceil \\ \leq & \overline{X}_t^{2,\hat{g}} \leq \max_i \left(\overline{X}_t^{i,\hat{g}} \right). \end{aligned} \tag{B.29}$$

Therefore, (B.18) follows from (B.27)-(B.29).

Proof of Claim 2

We set

$$Y_{t+1}^i := \left(Y_t^i - \tilde{D}_t^i \right)^+ + \tilde{A}_t^i \tag{B.30}$$

where Y_t^i satisfy the induction hypothesis, and $\tilde{A}_t^i, \tilde{D}_t^i, i = 1, 2$ are specified as follows. Let

$$M_x = \text{argmax}_i \{X_t^{i,\hat{g}}\}, \quad m_x = \text{argmin}_i \{X_t^{i,\hat{g}}\} \tag{B.31}$$

$$M_y = \text{argmax}_i \{Y_t^i\}, \quad m_y = \text{argmin}_i \{Y_t^i\}, \tag{B.32}$$

where $M_x = 1, m_x = 2$ (resp. $M_y = 1, m_y = 2$) when $\{X_t^{1,\hat{g}} = X_t^{2,\hat{g}}\}$ (resp. $\{Y_t^1 =$

$Y_t^2\}$); define

$$\left(\tilde{A}_t^{M_y}, \tilde{D}_t^{M_y}, \tilde{A}_t^{m_y}, \tilde{D}_t^{m_y}\right) := \begin{cases} \left(A_t^{M_x}, D_t^{m_x}, A_t^{m_x}, D_t^{M_x}\right) & \text{in case 1,} \\ \left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) & \text{in case 2,} \end{cases} \tag{B.33}$$

where the two cases are :

Case 1: $\{Y_t^{M_y} - 1 = X_t^{M_x,\hat{g}} = X_t^{m_x,\hat{g}}$ and $\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) = (0,1,1,0)$ or $(0,0,1,1)\}$.

Case 2: All other instances.

Assertion: The random variables Y_{t+1}^1, Y_{t+1}^2, defined by (B.30)-(B.33) satisfy (B.19)-(B.21).

As the proof of this assertion is long, we first provide a sketch of its proof and then we provide a full proof.

Sketch of the proof of the assertion

- Equation (B.33) implies the following: In case 2 we associate the arrival to and the departure from the longer queue M_x to those of the longer queue M_y, i.e. we set $\tilde{A}_t^{M_y} = A_t^{M_x}, \tilde{D}_t^{M_y} = D_t^{M_x}$. We do the same for the shorter queue m_x, m_y, i.e. $\tilde{A}_t^{m_y} = A_t^{m_x}, \tilde{D}_t^{m_y} = D_t^{m_x}$.

  In case 1, we have the same association for the arrivals as in case 2, that is $\tilde{A}_t^{M_y} = A_t^{M_x}, \tilde{A}_t^{m_y} = A_t^{m_x}$, but we reverse the association of the departures, that is $\tilde{D}_t^{M_y} = D_t^{m_x}, \tilde{D}_t^{m_y} = D_t^{M_x}$. Therefore the arrivals $\tilde{A}_t^i$, and departures $\tilde{D}_t^i$, have the same distribution as the original A_t^i, D_t^i, respectively, $i = 1, 2$. Then (B.19) follows from (B.30).

- To establish (B.20), we note that, because of (B.33), the sum of arrivals to (respectively, departures from) queues M_y and m_y equals to the sum of arrivals to (respectively, departures from) queues M_x and m_x.

  When $X_t^{i,\hat{g}}, Y_t^i \neq 0$, $i = 1, 2$, the function $(x-d)^+ + a$ is linear x, as $(x-d)^+ + a =$

$x - d + a$. Then from (B.30), (B.33) and the induction hypothesis we obtain

$$\begin{aligned} & Y_{t+1}^1 + Y_{t+1}^2 - \overline{X}_t^{1,\hat{g}} - \overline{X}_t^{2,\hat{g}} \\ =& Y_t^1 + Y_t^2 - X_t^{1,\hat{g}} - X_t^{2,\hat{g}} \geq 0 \end{aligned} \tag{B.34}$$

and this establish (B.20) when $X_t^{i,\hat{g}}, Y_t^i \neq 0$, $i = 1, 2$. In the full proof of the assertion, we show that show that (B.20) is also true when $X_t^{i,\hat{g}}, Y_t^i$ are not all non-zero.

- To establish (B.21) we consider the maximum of the queue lengths. In case 2, we show that (B.30)-(B.33) ensure that

$$Y_{t+1}^{M_y} \geq \overline{X}_t^{M_x,\hat{g}}, \tag{B.35}$$

$$\max\left(Y_{t+1}^{M_y}, Y_{t+1}^{m_y}\right) \geq \overline{X}_t^{m_x,\hat{g}}; \tag{B.36}$$

then (B.21) follows from (B.35)-(B.36).

In case 1 (B.21) is verified by direct computation in the full proof.

Proof of the assertion

For all $y_{0:t}$, we denote by $E_{y_{0:t}}$ the event $\{Y_{0:t}^i = y_{0:t}\}$.

Let $\tilde{Z}_t = \left(\tilde{A}_t^{M_y}, \tilde{D}_t^{M_y}, \tilde{A}_t^{m_y}, \tilde{D}_t^{m_y}\right)$, then for any realization $z_t \in \{0,1\}^4$ of $\tilde{Z}_t$ we have

$$\begin{aligned} & \mathbb{P}\left(\tilde{Z}_t = z_t | E_{y_{0:t}}\right) \\ =& \mathbb{P}\left(\tilde{Z}_t = z_t, \text{case } 1 | E_{y_{0:t}}\right) + \mathbb{P}\left(\tilde{Z}_t = z_t, \text{case } 2 | E_{y_{0:t}}\right). \end{aligned} \tag{B.37}$$

When $z_t \neq (0,1,1,0)$ or $(0,0,1,1)$, we get

$$\mathbb{P}\left(\tilde{Z}_t = z_t, \text{case } 1 | E_{y_{0:t}}\right) = 0, \tag{B.38}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\tilde{Z}_t = z_t, \text{case } 2 | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) = z_t | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^1, D_t^1, A_t^2, D_t^2\right) = z_t\right), \qquad \text{(B.39)}
\end{aligned}$$

where the last equality in (B.39) holds because the random variables $A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}$ are independent of $Y_0, Y_1, \ldots, Y_t$ and have the same distribution as $A_t^1, D_t^1, A_t^2, D_t^2$. Therefore, combining (B.38) and (B.39) we obtain for $z_t \neq (0,1,1,0)$ or $(0,0,1,1)$

$$\mathbb{P}\left(\tilde{Z}_t = z_t | E_{y_{0:t}}\right) = \mathbb{P}\left(\left(A_t^1, D_t^1, A_t^2, D_t^2\right) = z_t\right) \qquad \text{(B.40)}$$

When $z_t = (0,1,1,0)$ or $(0,0,1,1)$, let E denote the event $\{Y_t^{M_y} - 1 = X_t^{M_x,\hat{g}} = X_t^{m_x,\hat{g}}\}$; then we obtain

$$\begin{aligned}
&\mathbb{P}\left(\tilde{Z}_t = z_t, \text{case } 1 | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^{M_x}, D_t^{m_x}, A_t^{m_x}, D_t^{M_x}\right) = z_t, E | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^1, D_t^2, A_t^2, D_t^1\right) = z_t\right) \mathbb{P}\left(E | E_{y_{0:t}}\right), \qquad \text{(B.41)}
\end{aligned}$$

and

$$\begin{aligned}
&\mathbb{P}\left(\tilde{Z}_t = z_t, \text{case } 2 | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) = z_t, E^c | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^1, D_t^1, A_t^2, D_t^2\right) = z_t\right) \mathbb{P}\left(E^c | E_{y_{0:t}}\right), \qquad \text{(B.42)}
\end{aligned}$$

where the last equality in (B.41) and (B.42) follow by the fact that the random variables $A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}$ are independent of $Y_0, Y_1, \ldots, Y_t$ (hence, the event E which is generated by $Y_0, Y_1, \ldots, Y_t$) and have the same distribution as $A_t^1, D_t^1, A_t^2, D_t^2$.

Therefore, combining (B.41) and (B.42) we obtain for $z_t = (0,1,1,0)$ or $(0,0,1,1)$

$$\begin{aligned}
&\mathbb{P}\left(\tilde{Z}_t = z_t | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^1, D_t^2, A_t^2, D_t^1\right) = z_t\right) \mathbb{P}\left(E | E_{y_{0:t}}\right) \\
&+ \mathbb{P}\left(\left(A_t^1, D_t^1, A_t^2, D_t^2\right) = z_t\right) \mathbb{P}\left(E^c | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(A_t^1, D_t^1, A_t^2, D_t^2\right) = z_t\right), \qquad \text{(B.43)}
\end{aligned}$$

where the last equality in (B.43) is true because $A_t^1, D_t^1, A_t^2, D_t^2$ are independent and D_t^1 has the same distribution as D_t^2.

As a result of (B.40) and (B.43), for any $z_t \in \{0,1\}^4$ we have

$$\mathbb{P}\left(\tilde{Z}_t = z_t | E_{y_{0:t}}\right) = \mathbb{P}\left(\left(A_t^1, D_t^1, A_t^2, D_t^2\right) = z_t\right). \qquad \text{(B.44)}$$

Now consider any $y_{0:t+1}$. By (B.44) we have for $i = M_y$ or m_y

$$\begin{aligned}
&\mathbb{P}\left(Y_{t+1}^i = y_{t+1} | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(y_t^i - \tilde{D}_t^i\right)^+ + \tilde{A}_t^i = y_{t+1} | E_{y_{0:t}}\right) \\
=&\mathbb{P}\left(\left(y_t^i - D_t^i\right)^+ + A_t^i = y_{t+1}\right) \\
=&\mathbb{P}\left(X_{t+1}^{i,g0} = y_{t+1} | X_{0:t}^{i,g_0} = y_{0:t}\right). \qquad \text{(B.45)}
\end{aligned}$$

which is (B.19).

Now consider the sum $Y_{t+1}^1 + Y_{t+1}^2$.

From (B.33), we know that

$$\tilde{A}_t^{M_y} + \tilde{A}_t^{m_y} = A_t^{M_x} + A_t^{m_x} \quad a.s., \qquad \text{(B.46)}$$

$$\tilde{D}_t^{M_y} + \tilde{D}_t^{m_y} = D_t^{M_x} + D_t^{m_x} \quad a.s. \qquad \text{(B.47)}$$

Therefore, (B.46) implies

$$\begin{aligned} & Y_{t+1}^1 + Y_{t+1}^2 - \overline{X}_{t+1}^{1,\hat{g}} - \overline{X}_{t+1}^{1,\hat{g}} \\ = & \left(Y_t^{M_y} - \tilde{D}_t^{M_y}\right)^+ + \left(Y_t^{m_y} - \tilde{D}_t^{m_y}\right)^+ \\ & - \left(X_t^{M_x,\hat{g}} - D_t^{M_x}\right)^+ - \left(X_t^{m_x,\hat{g}} - D_t^{m_x}\right)^+. \end{aligned} \tag{B.48}$$

We proceed to show that the right hand side of (B.48) is positive. From the induction hypothesis for (3.69)-(3.68) we have

$$Y_t^{m_y} + Y_t^{M_y} \geq X_t^{m_x,\hat{g}} + X_t^{M_x,\hat{g}} \quad a.s., \tag{B.49}$$

$$Y_t^{M_y} \geq X_t^{M_x,\hat{g}} \quad a.s. \tag{B.50}$$

There are three possibilities: $\{Y_t^{M_y} = X_t^{M_x,\hat{g}}\}$, $\{Y_t^{M_y} > X_t^{M_x,\hat{g}}, X_t^{m_x,\hat{g}} = 0\}$ and $\{Y_t^{M_y} > X_t^{M_x,\hat{g}}, X_t^{m_x} > 0\}$.

First consider $\{Y_t^{M_y} = X_t^{M_x,\hat{g}}\}$. By (B.49) we have

$$Y_t^{m_y} \geq X_t^{m_x,\hat{g}} \quad a.s. \tag{B.51}$$

Note that $\{Y_t^{M_y} = X_t^{M_x,\hat{g}}\}$ belongs to case 2 in (B.33). From case 2 of (B.33) we also know that

$$D_t^{M_x} = \tilde{D}_t^{M_y}, \quad D_t^{m_x} = \tilde{D}_t^{m_y}. \tag{B.52}$$

Then, because of (B.50)-(B.52) we get

$$\begin{aligned} & \left(X_t^{M_x,\hat{g}} - D_t^{M_x}\right)^+ + \left(X_t^{m_x,\hat{g}} - D_t^{m_x}\right)^+ \\ \leq & \left(Y_t^{M_y} - D_t^{M_x}\right)^+ + \left(Y_t^{m_y} - D_t^{m_x}\right)^+ \\ = & \left(Y_t^{M_y} - \tilde{D}_t^{M_y}\right)^+ + \left(Y_t^{m_y} - \tilde{D}_t^{m_y}\right)^+ \quad a.s. \end{aligned} \tag{B.53}$$

If $Y_t^{M_y} > X_t^{M_x,\hat{g}}$ and $X_t^{m_x,\hat{g}} = 0$

$$\begin{aligned}
&\left(X_t^{M_x,\hat{g}} - D_t^{M_x}\right)^+ + \left(X_t^{m_x,\hat{g}} - D_t^{m_x}\right)^+ \\
=&\left(X_t^{M_x,\hat{g}} - D_t^{M_x}\right)^+ \\
\leq& X_t^{M_x,\hat{g}} \leq Y_t^{M_y} - 1 \\
\leq& \left(Y_t^{M_y} - \tilde{D}_t^{M_y}\right)^+ + \left(Y_t^{m_y} - \tilde{D}_t^{m_y}\right)^+
\end{aligned} \tag{B.54}$$

If $Y_t^{M_y} > X_t^{M_x,\hat{g}}$ and $X_t^{m_x} > 0$, then

$$\begin{aligned}
&\left(X_t^{M_x,\hat{g}} - D_t^{M_x}\right)^+ + \left(X_t^{m_x,\hat{g}} - D_t^{m_x}\right)^+ \\
=& X_t^{M_x,\hat{g}} - D_t^{M_x} + X_t^{m_x,\hat{g}} - D_t^{m_x} \\
=& X_t^{M_x,\hat{g}} + X_t^{m_x,\hat{g}} - \tilde{D}_t^{M_y} - \tilde{D}_t^{m_y} \\
\leq& Y_t^{M_y} + Y_t^{m_y} - \tilde{D}_t^{M_y} - \tilde{D}_t^{m_y} \\
\leq& \left(Y_t^{M_y} - \tilde{D}_t^{M_y}\right)^+ + \left(Y_t^{m_y} - \tilde{D}_t^{m_y}\right)^+
\end{aligned} \tag{B.55}$$

where the second equality in (B.55) follows from (B.47) and the first inequality in (B.55) follows from the induction hypothesis for (3.68).

The above results, namely (B.53)-(B.55), show that the right hand side of (B.48) is positive, and the proof for (B.20) is complete.

It remains to show that (B.21) is true.

We first consider case 2.

In case 2, we know from (B.33) that

$$\left(\tilde{A}_t^{M_y}, \tilde{D}_t^{M_y}, \tilde{A}_t^{m_y}, \tilde{D}_t^{m_y}\right) = \left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right). \tag{B.56}$$

Then,

$$\begin{aligned}\overline{X}_t^{M_x,\hat{g}} &= \left(X_t^{M_x,\hat{g}} - D_t^{M_x}\right)^+ + A_t^{M_x} \\ &= \left(X_t^{M_x,\hat{g}} - \tilde{D}_t^{M_y}\right)^+ + \tilde{A}_t^{M_y} \\ &\leq \left(Y_t^{M_y} - \tilde{D}_t^{M_y}\right)^+ + \tilde{A}_t^{M_y} \\ &= Y_{t+1}^{M_y}, \end{aligned} \tag{B.57}$$

where the second equality is a consequence of (B.56) and the inequality follows from the induction hypothesis for (3.69).

To proceed further we note that in case 2 there are three possibilities: $\{Y_t^{M_y} = X_t^{M_x,\hat{g}}\}$, $\{Y_t^{M_y} - 2 \geq X_t^{m_x,\hat{g}}\}$ and $\{Y_t^{M_y} > X_t^{M_x,\hat{g}}, Y_t^{M_y} - 2 < X_t^{m_x,\hat{g}}\}$

If $Y_t^{M_y} = X_t^{M_x,\hat{g}}$, (B.51) is also true. Following similar arguments as in (B.57) we obtain

$$\overline{X}_t^{m_x,\hat{g}} \leq Y_{t+1}^{m_y}. \tag{B.58}$$

If $Y_t^{M_y} - 2 \geq X_t^{m_x,\hat{g}}$

$$\overline{X}_t^{m_x,\hat{g}} \leq X_t^{m_x,\hat{g}} + 1 \leq Y_t^{M_y} - 1 \leq Y_{t+1}^{M_y}. \tag{B.59}$$

If $Y_t^{M_y} > X_t^{M_x,\hat{g}}$ and $Y_t^{M_y} - 2 < X_t^{m_x,\hat{g}}$ it can only be $Y_t^{M_y} - 1 = X_t^{M_x,\hat{g}} = X_t^{m_x,\hat{g}}$. Since we are in case 2, $\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) \neq (0,1,1,0)$. Therefore,

$$A_t^{m_x} - D_t^{m_x} \leq A_t^{M_x} - D_t^{M_x} + 1. \tag{B.60}$$

Then we get

$$\begin{aligned}
\overline{X}_t^{m_x,\hat{g}} &= \left(Y_t^{M_y} - 1 - D_t^{m_x}\right)^+ + A_t^{m_x} \\
&= \max\left(A_t^{m_x}, Y_t^{M_y} - 1 - D_t^{m_x} + A_t^{m_x}\right) \\
&\leq \max\left(A_t^{m_x}, Y_t^{M_y} - D_t^{M_x} + A_t^{M_x}\right) \\
&\leq \max\left(A_t^{m_x}, Y_{t+1}^{M_y}\right) \\
&\leq \max\left(Y_{t+1}^{m_y}, Y_{t+1}^{M_y}\right). \qquad \text{(B.61)}
\end{aligned}$$

Combining (B.57), (B.58), (B.59) and (B.61) we get (B.21) when case 2 is true.

Now consider case 1. We have $Y_t^{M_y} - 1 = X_t^{M_x,\hat{g}} = X_t^{m_x,\hat{g}}$.

When $\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) = (0, 1, 1, 0)$, then

$$\begin{aligned}
\overline{X}_t^{M_x,\hat{g}} &= \left(X_t^{M_x,\hat{g}} - 1\right)^+ \\
&\leq \overline{X}_t^{m_x} \\
&= X_t^{m_x} + 1 \\
&= \left(Y_t^{M_y} - D_t^{m_x}\right)^+ + A_t^{M_x} \\
&= Y_{t+1}^{M_y} \qquad \text{(B.62)}
\end{aligned}$$

When $\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) = (0,0,1,1)$ we get

$$\begin{aligned}
\overline{X}_t^{M_x,\hat{g}} =& X_t^{M_x,\hat{g}} \\
\leq& \overline{X}_t^{m_x,\hat{g}} \\
=& \max\left(X_t^{m_x,\hat{g}}, 1\right) \\
=& \max\left(\left(Y_t^{M_y} - D_t^{m_x}\right)^+ + A_t^{M_x}, A_t^{m_x}\right) \\
=& \max\left(Y_{t+1}^{M_y}, A_t^{m_x}\right) \\
\leq& \max\left(Y_{t+1}^{M_y}, Y_{t+1}^{m_y}\right). \qquad \text{(B.63)}
\end{aligned}$$

Combining (B.62) and (B.63) we obtain (B.21) for case 1.

As a result, (B.21) holds for both cases 1 and 2.

Remark:

We note that we need the two cases described in (B.33) for the following reasons. If we eliminate case 1 and always associate $\left(\tilde{A}_t^{M_y}, \tilde{D}_t^{M_y}, \tilde{A}_t^{m_y}, \tilde{D}_t^{m_y}\right)$ with $\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right)$ as in case 2, then when $\{Y_t^{M_y} - 1 = X_t^{m_x,\hat{g}}$ and $\left(A_t^{M_x}, D_t^{M_x}, A_t^{m_x}, D_t^{m_x}\right) = (0,1,1,0)\}$, the shorter queue m_x increases by one customer, and the longer queue M_y decreases by one customer; therefore $\overline{X}_t^{m_x,\hat{g}} = Y_{t+1}^{M_y} + 1$ and (B.21) is not satisfied.

□

Proof of Lemma III.16. From Lemma III.15, at any time t there exists Y_t^i such that such that (3.67)-(3.69) hold.

Adopting the notations M_x, m_x and M_y, m_y in the proof of Lemma III.15, we have at every time t

$$X_t^{m_x,\hat{g}} \leq X_t^{M_X,\hat{g}} \quad a.s., \qquad \text{(B.64)}$$

$$Y_t^{m_y} \leq Y_t^{M_y} \quad a.s. \qquad \text{(B.65)}$$

Furthermore, from (3.69) we have

$$X_t^{M_x,\hat{g}} \leq Y_t^{M_y} \quad a.s. \tag{B.66}$$

If $X_t^{m_x,\hat{g}} \leq Y_t^{m_y}$, (B.66) and the fact that $c(\cdot)$ is increasing give

$$c\left(X_t^{M_X,\hat{g}}\right) + c\left(X_t^{m_X,\hat{g}}\right) \leq c\left(Y_t^{M_y}\right) + c\left(Y_t^{m_y}\right). \tag{B.67}$$

If $X_t^{m_x,\hat{g}} > Y_t^{m_y}$, then

$$Y_t^{m_y} < X_t^{m_x,\hat{g}} \leq X_t^{M_x,\hat{g}} \leq Y_t^{M_y}. \tag{B.68}$$

Since $c(\cdot)$ is convex, it follows from (B.68) that

$$\frac{c\left(Y_t^{M_y}\right) - c\left(X_t^{M_x,\hat{g}}\right)}{Y_t^{M_y} - X_t^{M_x,\hat{g}}} \geq \frac{c\left(X_t^{m_x,\hat{g}}\right) - c\left(Y_t^{m_y}\right)}{X_t^{m_x,\hat{g}} - Y_t^{m_y}}. \tag{B.69}$$

From (3.68) in Lemma III.15 we know that

$$Y_t^{M_y} - X_t^{M_x,\hat{g}} \geq X_t^{m_x,\hat{g}} - Y_t^{m_y}. \tag{B.70}$$

Combining (B.69) and (B.70) we get

$$c\left(Y_t^{M_y}\right) + c\left(Y_t^{m_y}\right) \geq c\left(X_t^{M_x,\hat{g}}\right) + c\left(X_t^{m_x,\hat{g}}\right). \tag{B.71}$$

□

Proof of Lemma III.17. Let $\{Y_t^1, t \in \mathbb{Z}_+\}$ and $\{Y_t^2, t \in \mathbb{Z}_+\}$ be the processes defined in Lemma III.15. Then $\{Y_t^i, t \in \mathbb{Z}_+\}$ has the same distribution as $\{X_t^{i,g_0}, t \in \mathbb{Z}_+\}$ for $i = 1, 2$.

Since $\mu > \lambda$, the processes $\{Y_t^i, t \in \mathbb{Z}_+\}, i = 1, 2$ are irreducible positive recurrent

Markov chains. Moreover, the two processes $\{Y_t^1, t \in \mathbb{Z}_+\}$ and $\{Y_t^2, t \in \mathbb{Z}_+\}$ have the same stationary distribution, denoted by π^{g_0}. Under Assumption III.11, by Ergodic theorem of Markov chains (see [69, chap. 3]) we get

$$\begin{aligned}\lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(Y_t^1) &= \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} c(Y_t^2) \\ &= \sum_{x=0}^{\infty} \pi^{g_0}(x) c(x) \quad a.s. \end{aligned} \tag{B.72}$$

Let $W_T^i(Y_{0:T-1}) := \frac{1}{T} \sum_{t=0}^{T-1} c(Y_t^i), i = 1, 2.$

We show that $\{W_T^i(Y_{0:T-1}), T = 1, 2, \dots\}$ is uniformly integrable for $i = 1, 2$. That is,

$$\sup_T \mathbb{E}\left[W_T^i(Y_{0:T-1}) 1_{\{W_T^i(Y_{0:T-1})>N\}}\right] \to 0 \tag{B.73}$$

as $N \to \infty$.

Let $p^{g_0}(x, y), x, y \in \mathbb{Z}_+$ be the transition probabilities of the Markov chain. Note that the initial PMF of the process $\{Y_t^i, t \in \mathbb{Z}_+\}, i = 1, 2$ is π_0^i. From Assumption III.11 we know that $\pi_0^i(x) = 0, i = 1, 2$ for all $x > M$.

Letting $R := \max_{x \le M} \frac{\pi_0^i(x)}{\pi^{g_0}(x)} < \infty$, we obtain for $i = 1, 2$

$$\begin{aligned}&\mathbb{E}\left[W_T^i(Y_{0:T-1}) 1_{\{W_T^i(Y_{0:T-1})>N\}}\right] \\ =& \sum_{y_{0:T-1}} W_T^i(y_{0:T-1}) 1_{\{W_T^i(y_{0:T-1})>N\}} \mathbb{P}(Y_{0:T-1} = y_{0:T-1}) \\ =& \sum_{y_{0:T-1}} W_T^i(y_{0:T-1}) 1_{\{W_T^i(y_{0:T-1})>N\}} \pi_0^i(y_0) \Pi_{t=1}^{T-1} p^{g_0}(y_{t-1}, y_t) \\ \le& R \sum_{y_{0:T-1}} W_T^i(y_{0:T-1}) 1_{\{W_T^i(y_{0:T-1})>N\}} \pi^{g_0}(y_0) \Pi_{t=1}^{T-1} p^{g_0}(y_{t-1}, y_t) \\ =& R\, \mathbb{E}\left[W_T^{\pi^{g_0}} 1_{\{W_T^{\pi^{g_0}}>N\}}\right], \end{aligned} \tag{B.74}$$

where $W_T^{\pi^{g_0}} = \frac{1}{T}\sum_{t=0}^{T-1} c(Y_t^{\pi^{g_0}})$ and $\{Y_t^{\pi^{g_0}}, t \in \mathbb{Z}_+\}$ is the chain with transition probabilities $p^{g_0}(x,y)$ and initial PMF π^{g_0}.

Note that $\{Y_t^{\pi^{g_0}}, t \in \mathbb{Z}_+\}$ is stationary because the initial PMF is the stationary distribution π^{g_0}. From Birkhoff's Ergodic theorem we know that $\{W_T^{\pi^{g_0}}, T = 1, 2, \dots\}$ converges $a.s.$ and in expectation (see [105, chap. 2]). Therefore, $\{W_T^{\pi^{g_0}}, T = 1, 2, \dots\}$ is uniformly integrable, and the right hand side of (B.74) goes to zeros uniformly as $N \to \infty$. Consequently, $\{W_T^i(Y_{0:T-1}), T = 1, 2, \dots\}$ is also uniformly integrable for $i = 1, 2$.

Since $W_T = W_T^1(Y_{0:T-1}) + W_T^2(Y_{0:T-1})$ for all $T = 1, 2, \dots$, $\{W_T, T = 1, 2, \dots\}$ is uniformly integrable.

□

Proof of Corollary III.18. From Lemma III.16, there exists $\{Y_t^1, Y_t^2, t \in \mathbb{Z}_+\}$ such that (3.67) holds and

$$c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right) \leq c\left(Y_t^1\right) + c\left(Y_t^2\right) \quad a.s. \tag{B.75}$$

Let

$$W_T := \frac{1}{T}\sum_{t=0}^{T-1}\left(c\left(Y_t^1\right) + c\left(Y_t^2\right)\right), \tag{B.76}$$

$$V_T := \frac{1}{T}\sum_{t=0}^{T-1}\left(c\left(X_t^{1,\hat{g}}\right) + c\left(X_t^{2,\hat{g}}\right)\right). \tag{B.77}$$

From (B.75) it follows that

$$V_T \leq W_T, T = 1, 2, \dots \tag{B.78}$$

From Lemmas III.17, $\{W_T, T = 1, 2, \dots\}$ is uniformly integrable, therefore $\{V_T, T = 1, 2, \dots\}$, which is bounded above by $\{W_T, T = 1, 2, \dots\}$ is also uniformly integrable.

From the property of uniformly integrability, if $\{V_T, T = 1, 2, \dots\}$ converges a.s., we know that $\{V_T, T = 1, 2, \dots\}$ also converges in expectation. Furthermore,

$$\begin{aligned} J^{\hat{g}}\left(\pi_0^1, \pi_0^2\right) &= \limsup_{T\to\infty} \frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T-1}\left(c\left(Y_t^1\right) + c\left(Y_t^2\right)\right)\right] \\ &= \limsup_{T\to\infty} \mathbb{E}\left[V_T\right] \\ &\leq \limsup_{T\to\infty} \mathbb{E}\left[W_T\right] = J^{g_0}. \end{aligned} \tag{B.79}$$

□

Proof of Lemma III.19. First we show that $\{S_t, t \geq T_0 + 1\}$ is a Markov chain. For $s_t \geq 2$,

$$\begin{aligned} &\mathbb{P}\left(S_{t+1} = s_{t+1} | S_{T_0+1:t} = s_{T_0+1:t}\right) \\ =&\mathbb{P}\left(\left(s_t - D_t^1 - D_t^2 + A_t^1 + A_t^2\right) = s_{t+1} | S_{T_0+1:t} = s_{T_0+1:t}\right) \\ =&\mathbb{P}\left(\left(s_t - D_t^1 - D_t^2 + A_t^1 + A_t^2\right) = s_{t+1} | S_t = s_t\right) \\ =&\mathbb{P}\left(S_{t+1} = s_{t+1} | S_t = s_t\right). \end{aligned} \tag{B.80}$$

The first and last equalities in (B.80) follow from the construction of the process $\{S_t, t \geq T_0 + 1\}$. The second equality in (B.80) is true because T_0 is a stopping time with respect to $\{X_t^{1,\hat{g}}, X_t^{2,\hat{g}}, t \in \mathbb{Z}_+\}$, and $A_t^i, D_t^i, i = 1, 2$ are independent of all random variables before t. Similarly, for $s_t = 0$ we have, by arguments similar to the above,

$$\begin{aligned} &\mathbb{P}\left(S_{t+1} = s_{t+1} | S_{T_0+1:t} = s_{T_0+1:t}\right) \\ =&\mathbb{P}\left(A_t^1 + A_t^2 = s_{t+1} | S_{T_0+1:t-1} = s_{T_0+1:t-1}, S_t = 0\right) \\ =&\mathbb{P}\left(A_t^1 + A_t^2 = s_{t+1} | S_t = 0\right) \\ =&\mathbb{P}\left(S_{t+1} = s_{t+1} | S_t = 0\right). \end{aligned} \tag{B.81}$$

The first and last equality in (B.81) follow from the construction of the process $\{S_t, t \geq T_0 + 1\}$. The second equality in (B.81) is true because $A_t^i, D_t^i, i = 1, 2$ are independent of all variables before t. For $s_t = 1$,

$$\begin{aligned}
&\mathbb{P}\left(S_{t+1} = s_{t+1} | S_{T_0+1:t} = s_{T_0+1:t}\right) \\
=&\mathbb{P}\left(s_t + 1_{\left\{X_t^{1,\hat{g}}=0\right\}}(D_t^1 - D_t^2) - D_t^1 + A_t^1 + A_t^2 = s_{t+1} | S_{T_0+1:t} = s_{T_0+1:t}\right) \\
=&\mathbb{P}\left(1 - D_t^2 + A_t^1 + A_t^2 = s_{t+1}, X_t^{1,\hat{g}} = 0 | S_{T_0+1:t-1} = s_{T_0+1:t-1}, S_t = 1\right) \\
&+ \mathbb{P}\left(1 - D_t^1 + A_t^1 + A_t^2 = s_{t+1}, X_t^{1,\hat{g}} = 1 | S_{T_0+1:t-1} = s_{T_0+1:t-1}, S_t = 1\right) \\
=&\mathbb{P}\left(1 - D_t^1 + A_t^1 + A_t^2 = s_{t+1}, X_t^{1,\hat{g}} = 0 | S_{T_0+1:t-1} = s_{T_0+1:t-1}, S_t = 1\right) \\
&+ \mathbb{P}\left(1 - D_t^1 + A_t^1 + A_t^2 = s_{t+1}, X_t^{1,\hat{g}} = 1 | S_{T_0+1:t-1} = s_{T_0+1:t-1}, S_t = 1\right) \\
=&\mathbb{P}\left(1 - D_t^1 + A_t^1 + A_t^2 = s_{t+1} | S_{T_0+1:t-1} = s_{T_0+1:t-1}, S_t = 1\right) \\
=&\mathbb{P}\left(1 - D_t^1 + A_t^1 + A_t^2 = s_{t+1} | S_t = 1\right) \\
=&\mathbb{P}\left(S_{t+1} = s_{t+1} | S_t = s_t\right). \qquad \text{(B.82)}
\end{aligned}$$

The first equality in (B.82) follows from the construction of the process $\{S_t, t \geq T_0 + 1\}$. The second and forth equalities follow from the fact that $X_t^{1,\hat{g}}$ can be either 0 or 1. In the third equality, D_t^2 is replaced by D_t^1 in the first term; this is true because D_t^1 and D_t^2 are identically distributed and independent of $X_t^{1,\hat{g}}$ and all past random variables. The fifth equality holds because T_0 is a stopping time with respect to $\{X_t^{1,\hat{g}}, X_t^{2,\hat{g}}, t \in \mathbb{Z}_+\}$ and $A_t^i, D_t^i, i = 1, 2$ are independent of all past random variables. The last equality follows from the same arguments that lead to the first through the fifth equalities.

Therefore, the process $\{S_t, t \geq T_0 + 1\}$ is a Markov chain.

Since $\lambda, \mu > 0$, the Markov chain is irreducible.

We prove that the process $\{S_t, t \geq T_0 + 1\}$ is positive recurrent. Note that, for all

$s = 0, 1, 2, \ldots$, because of the construction of $\{S_t, t \geq T_0 + 1\}$

$$\begin{aligned}
&\mathbb{E}\left[S_{t+1} | S_t = s\right] \\
\leq &\mathbb{E}\left[S_t + A_t^1 + A_t^2 | S_t = s\right] \\
= &s + 2\lambda < \infty.
\end{aligned} \tag{B.83}$$

Moreover, for all $s \geq 2$,

$$\begin{aligned}
&\mathbb{E}\left[S_{t+1} | S_t = s\right] \\
= &\mathbb{E}\left[s - D_t^1 - D_t^2 + A_t^1 + A_t^2 | S_t = s\right] \\
= &s - 2\mu + 2\lambda < s.
\end{aligned} \tag{B.84}$$

Using Foster's theorem (see [69, chap. 5]), we conclude that the Markov chain $\{S_t, t \geq T_0 + 1\}$ is positive recurrent. □

Proof of Lemma III.20. Let $(\Omega, \mathcal{F}, \mathbb{P})$ denote the basic probability space for our problem. Define events $E_t \in \mathcal{F}, t = 0, 1, \ldots$ to be

$$E_t = \{\omega \in \Omega : \left(U_{t'}^{1,\hat{g}}(\omega), U_{t'}^{2,\hat{g}}(\omega)\right) \neq (0,0) \quad \forall t' \geq t\} \tag{B.85}$$

If the claim of this lemma is not true, we get

$$\mathbb{P}\left(\bigcup_{t=0}^{\infty} E_t\right) = 1 - \mathbb{P}\left(\left(U_t^{1,\hat{g}}, U_t^{2,\hat{g}}\right) = (0,0) \quad i.o.\right) > 0. \tag{B.86}$$

Therefore, there exist some t_0 such that $\mathbb{P}(E_{t_0}) > 0$. Since t_0 is a constant, it is a stopping time with respect to $\{X_t^{1,\hat{g}}, X_t^{2,\hat{g}}, t \in \mathbb{Z}_+\}$.

Consider the process $\{S_t, t = t_0 + 1, t_0 + 2, ...\}$ defined in Lemma III.19 with the stopping time t_0. From Lemma III.19 we know that $\{S_t, t \geq t_0 + 1\}$ is an irreducible positive recurrent Markov chain. Furthermore, along the sample path induced by any

$\omega \in E_{t_0}$, we claim that for all $t \geq t_0 + 1$

$$\begin{aligned} S_t(\omega) =& X_t^{1,\hat{g}}(\omega) + X_t^{2,\hat{g}}(\omega) \\ =& \overline{X}_{t-1}^{1,\hat{g}}(\omega) + \overline{X}_{t-1}^{2,\hat{g}}(\omega). \end{aligned} \tag{B.87}$$

The claim is shown by induction below.

By the definition of $\{S_t, t \geq t_0 + 1\}$ in Lemma III.19, we have at time $t_0 + 1$ for any $\omega \in E_{t_0}$

$$\begin{aligned} S_{t_0+1}(\omega) =& X_{t_0+1}^{1,\hat{g}}(\omega) + X_{t_0+1}^{2,\hat{g}}(\omega) \\ =& \overline{X}_{t_0}^{1,\hat{g}}(\omega) + \overline{X}_{t_0}^{2,\hat{g}}(\omega), \end{aligned} \tag{B.88}$$

where the last inequality in (B.88) follows from the system dynamics (3.1)-(3.3).

Assume equation (B.87) is true at time t ($t \geq t_0 + 1$). At time $t + 1$ we have, by (3.1)-(3.3),

$$\begin{aligned} & X_{t+1}^{1,\hat{g}} + X_{t+1}^{2,\hat{g}} \\ =& (X_t^{1,\hat{g}} - D_t^1)^+ + (X_t^{2,\hat{g}} - D_t^2)^+ + A_t^1 + A_t^2 \\ =& X_t^{1,\hat{g}} + X_t^{2,\hat{g}} - D_t^1 - D_t^2 + A_t^1 + A_t^2 \\ & + D_t^1 1_{\left\{X_t^{1,\hat{g}}=0\right\}} + D_t^2 1_{\left\{X_t^{2,\hat{g}}=0\right\}}. \end{aligned} \tag{B.89}$$

Since along the sample path induced by $\omega \in E_{t_0}$, $\left(U_{t-1}^{1,\hat{g}}(\omega), U_{t-1}^{2,\hat{g}}(\omega)\right) \neq (0,0)$ and $X_t^{i,\hat{g}} = \overline{X}_{t-1}^{i,\hat{g}} - U_{t-1}^{i,\hat{g}} + U_{t-1}^{j,\hat{g}}$, the event $\{X_t^{i,\hat{g}} = 0\} \bigcap E_{t_0}$ ($i = 1$ or 2) implies that $\overline{X}_{t-1}^{i,\hat{g}} = 1$, $U_{t-1}^{i,\hat{g}} = 1$ and $U_{t-1}^{j,\hat{g}} = 0$. For this case, $\overline{X}_{t-1}^{i,\hat{g}} = 1$ and $U_{t-1}^{i,\hat{g}} = 1$ further imply that the threshold is smaller than one. Then, the only possibility for $U_{t-1}^{j,\hat{g}} = 0$

is $\overline{X}_{t-1}^{j,\hat{g}} = 0$. Therefore,

$$\begin{aligned}
&\left\{X_t^{i,\hat{g}} = 0\right\} \bigcap E_{t_0} \\
\subseteq &\left\{\overline{X}_{t-1}^{i,\hat{g}} = 1,\, U_{t-1}^{i,\hat{g}} = 1,\, \overline{X}_{t-1}^{j,\hat{g}} = 0 \text{ and } U_{t-1}^{j,\hat{g}} = 0\right\} \\
\subseteq &\{S_t = 1\}. \qquad \text{(B.90)}
\end{aligned}$$

Consequently, from (B.90), for any $\omega \in E_{t_0}$

$$\begin{aligned}
&D_t^1(\omega) 1_{\left\{X_t^{1,\hat{g}}(\omega)=0\right\}} + D_t^2(\omega) 1_{\left\{X_t^{2,\hat{g}}(\omega)=0\right\}} \\
=&1_{\{S_t(\omega)=1\}} \left(D_t^1(\omega) 1_{\left\{X_t^{1,\hat{g}}(\omega)=0\right\}} + D_t^2(\omega) 1_{\left\{X_t^{2,\hat{g}}(\omega)=0\right\}}\right). \\
=&1_{\{S_t(\omega)=1\}} \left(1_{\left\{X_t^{1,\hat{g}}(\omega)=0\right\}} (D_t^1(\omega) - D_t^2(\omega)) + D_t^2(\omega)\right). \qquad \text{(B.91)}
\end{aligned}$$

Moreover, $\left(U_{t-1}^{1,\hat{g}}(\omega), U_{t-1}^{2,\hat{g}}(\omega)\right) \neq (0,0)$ implies that $\left(\overline{X}_{t-1}^{1,\hat{g}}(\omega), \overline{X}_{t-1}^{2,\hat{g}}(\omega)\right) \neq (0,0)$. Hence,

$$S_t(\omega) = \overline{X}_{t-1}^{1,\hat{g}}(\omega) + \overline{X}_{t-1}^{2,\hat{g}}(\omega) \neq 0, \qquad \text{(B.92)}$$

and

$$\begin{aligned}
&X_{t+1}^{1,\hat{g}}(\omega) + X_{t+1}^{2,\hat{g}}(\omega) \\
=&X_t^{1,\hat{g}}(\omega) + X_t^{2,\hat{g}}(\omega) - D_t^1(\omega) - D_t^2(\omega) + A_t^1(\omega) + A_t^2(\omega) \\
&+ 1_{\{S_t(\omega)=1\}} \left(1_{\left\{X_t^{1,\hat{g}}(\omega)=0\right\}} (D_t^1(\omega) - D_t^2(\omega)) + D_t^2(\omega)\right) \\
=&X_t^{1,\hat{g}}(\omega) + X_t^{2,\hat{g}}(\omega) - D_t^1(\omega) - D_t^2(\omega) + A_t^1(\omega) + A_t^2(\omega) \\
&+ 1_{\{S_t(\omega)=1\}} \left(1_{\left\{X_t^{1,\hat{g}}(\omega)=0\right\}} (D_t^1(\omega) - D_t^2(\omega)) + D_t^2(\omega)\right) \\
&+ 1_{\{S_t(\omega)=0\}} \left(D_t^1(\omega) + D_t^2(\omega)\right) \\
=&S_{t+1}(\omega), \qquad \text{(B.93)}
\end{aligned}$$

where the first and second equalities in (B.93) follow from (B.91) and (B.92), respectively. The last equality in (B.93) follows from the construction of $\{S_t, t \geq t_0 + 1\}$. Furthermore, by the system dynamics (3.1)-(3.3) we have

$$\begin{aligned}\overline{X}_t^{1,\hat{g}}(\omega) + \overline{X}_t^{2,\hat{g}}(\omega) =& X_{t+1}^{1,\hat{g}}(\omega) + X_{t+1}^{2,\hat{g}}(\omega) \\ =& S_{t+1}(\omega). \end{aligned} \tag{B.94}$$

Thus, equation (B.87) is true for any $\omega \in E_{t_0}$ for all $t \geq t_0 + 1$.

Then, for any $\omega \in E_{t_0}$

$$S_t(\omega) = \overline{X}_{t-1}^{1,\hat{g}}(\omega) + \overline{X}_{t-1}^{2,\hat{g}}(\omega) \neq 0 \text{ for all } t \geq t_0 + 1 \tag{B.95}$$

because $\left(U_{t-1}^{1,\hat{g}}(\omega), U_{t-1}^{2,\hat{g}}(\omega)\right) \neq (0,0)$ for all $t \geq t_0 + 1$. Since $\mathbb{P}(E_{t_0}) > 0$, (B.95) contradicts the fact that $\{S_t, t \geq t_0 + 1\}$ is recurrent.

Therefore, no such event $E_{t_0} \in \mathcal{F}$ with positive probability exists, and the proof of this lemma is complete.

□

Proof of Lemma III.22. For any fixed centralized policy $g \in \mathcal{G}_c$, the information I_t^1, I_t^2 available to the centralized controller includes all primitive random variables $X_0^i, A_{0:t}^i, D_{0:t}^i, i = 1, 2$ up to time t. Since all other random variables are functions of these primitive random variables and g, we have

$$\begin{aligned} U_t^{i,g} =& g_t^i(I_t^1, I_t^2) \\ =& g_t^i(X_0^1, X_0^2, A_{0:t}^1, A_{0:t}^2, D_{0:t}^1, D_{0:t}^2), \end{aligned} \tag{B.96}$$

for $i = 1, 2$. For any initial queue lengths x_0^1, x_0^2, we now define a policy $\tilde{g}$ from g for the case when both queues are initially empty. Let $\tilde{g}$ be the policy such that for

$i = 1, 2$

$$\begin{aligned} U_t^{i,\tilde{g}} =& \tilde{g}_t^i(I_t^1, I_t^2) \\ :=& \begin{cases} g_t^i(x_0^1, x_0^2, A_{0:t}^1, A_{0:t}^2, D_{0:t}^i, D_{0:t}^2) & \text{if } \overline{X}_t^{i,\tilde{g}} > 0 \\ 0 & \text{if } \overline{X}_t^{i,\tilde{g}} = 0 \end{cases} \\ =& \min\left(U_t^{i,g}, \overline{X}_t^{i,\tilde{g}}\right) \leq U_t^{i,g}, \end{aligned} \tag{B.97}$$

where $X_t^{1,\tilde{g}}$ and $X_t^{2,\tilde{g}}$ denote the queue lengths at time t due to policy $\tilde{g}$ with initial queue lengths $X_0^{1,\tilde{g}} = X_0^{2,\tilde{g}} = 0$.

At time 0 we have $X_0^{i,g} = x_0^i \geq 0 = X_0^{i,\tilde{g}}$ for $i = 1, 2$. We now prove by induction that for all time t

$$X_t^{i,g} \geq X_t^{i,\tilde{g}}, \quad i = 1, 2. \tag{B.98}$$

Suppose the claim is true at time t. Then, from the system dynamics (3.1)-(3.2) and (B.98) we obtain, for $i = 1, 2$,

$$\begin{aligned} \overline{X}_t^{i,g} &= \left(X_t^{i,g} - D_t^i\right)^+ + A_t^i \\ &\geq \left(X_t^{i,\tilde{g}} - D_t^i\right)^+ + A_t^i = \overline{X}_t^{i,\tilde{g}}. \end{aligned} \tag{B.99}$$

Furthermore from (3.1)-(3.2) and (B.97)

$$\begin{aligned} X_{t+1}^{i,g} =& \overline{X}_t^{i,g} - U_t^{i,g} + U_t^{j,g} \\ \geq& \overline{X}_t^{i,g} - U_t^{i,g} + U_t^{j,\tilde{g}} \end{aligned} \tag{B.100}$$

If $\overline{X}_t^{i,\tilde{g}} > 0$, then, because of (B.97) and (B.99)

$$\begin{aligned}\overline{X}_t^{i,g} - U_t^{i,g} =& \overline{X}_t^{i,g} - \min\left(U_t^{i,g}, \overline{X}_t^{i,\tilde{g}}\right) \\ =& \overline{X}_t^{i,g} - U_t^{i,\tilde{g}} \geq \overline{X}_t^{i,\tilde{g}} - U_t^{i,\tilde{g}}. \end{aligned} \tag{B.101}$$

If $\overline{X}_t^{i,\tilde{g}} = 0$, since $\overline{X}_t^{i,g} - U_t^{i,g} \geq 0$, (B.97) implies

$$\overline{X}_t^{i,g} - U_t^{i,g} \geq 0 = \overline{X}_t^{i,\tilde{g}} - U_t^{i,\tilde{g}}. \tag{B.102}$$

Combining (B.100)-(B.102) and (3.1)-(3.2) we get

$$\begin{aligned}X_{t+1}^{i,g} \geq& \overline{X}_t^{i,g} - U_t^{i,g} + U_t^{j,\tilde{g}} \\ \geq& \overline{X}_t^{i,\tilde{g}} - U_t^{i,\tilde{g}} + U_t^{j,\tilde{g}} = X_{t+1}^{i,\tilde{g}}. \end{aligned} \tag{B.103}$$

Therefore, we complete the proof of the claim (B.98).

Since the cost function is increasing, (B.98) implies that for all $g \in \mathcal{G}_c$ and any initial condition $X_0^1 = x_0^1, X_0^2 = x_0^2$,

$$\inf_{g \in \mathcal{G}_c} J_T^g(0,0) \leq J_T^{\tilde{g}}(0,0) \leq J_T^g(x_0^1, x_0^2). \tag{B.104}$$

Consequently, for any PMFs π_0^1, π_0^2

$$\inf_{g \in \mathcal{G}_c} J_T^g(0,0) \leq \inf_{g \in \mathcal{G}_c} J_T^g(\pi_0^1, \pi_0^2). \tag{B.105}$$

Moreover, the result of Lemma III.9 ensures that $\hat{g}$ gives the smallest expected cost among policies in $\mathcal{G}_c$ for any finite horizon when $X_0^1 = X_0^2 = 0$. It follows that, for

any finite T,

$$J_T^{\hat{g}}(0,0) = \inf_{g \in \mathcal{G}_c} J_T^g(0,0) \leq J_T^{\tilde{g}}(0,0) \leq J_T^g(x_0^1, x_0^2). \tag{B.106}$$

For infinite horizon cost, we divide each term in (B.106) by T and let T to infinity, and we obtain, for any π_0^1, π_0^2,

$$J^{\hat{g}}(0,0) = \inf_{g \in \mathcal{G}_c} J^g(0,0) \leq J^{\tilde{g}}(0,0) \leq J^g(x_0^1, x_0^2). \tag{B.107}$$

□

Proof of the claim in the proof of Theorem III.14. We prove here our claim expressed by equation (3.90) to complete the proof of Theorem III.14. By (3.78),

$$S_{T_0+1} = X_{T_0+1}^{1,\hat{g}} + X_{T_0+1}^{2,\hat{g}}. \tag{B.108}$$

We prove by induction that $X_t^{1,\hat{g}} + X_t^{2,\hat{g}} = S_t$ for all $t \geq T_0 + 1$.
Assume that $X_t^{1,\hat{g}} + X_t^{2,\hat{g}} = S_t$ at time t, $t \geq T_0 + 1$. Then for time $t+1$, because of the systems dynamics (3.1)-(3.3),

$$\begin{aligned}
&X_{t+1}^{1,\hat{g}} + X_{t+1}^{2,\hat{g}} \\
=&(X_t^{1,\hat{g}} - D_t^1)^+ + (X_t^{2,\hat{g}} - D_t^2)^+ + A_t^1 + A_t^2 \\
=&X_t^{1,\hat{g}} + X_t^{2,\hat{g}} - D_t^1 - D_t^2 + A_t^1 + A_t^2 \\
&+ D_t^1 1_{\left\{X_t^{1,\hat{g}}=0\right\}} + D_t^2 1_{\left\{X_t^{2,\hat{g}}=0\right\}}.
\end{aligned} \tag{B.109}$$

When $X_t^{i,\hat{g}} = 0$ ($i = 1$ or 2), $U_{t-1}^{j,\hat{g}}$ should be 0 because

$$0 = X_t^{i,\hat{g}} = \overline{X}_{t-1}^{i,\hat{g}} - U_{t-1}^{i,\hat{g}} + U_{t-1}^{j,\hat{g}} \tag{B.110}$$

and $\overline{X}_{t-1}^{i,\hat{g}} - U_{t-1}^{i,\hat{g}} \geq 0$.

We consider the following two cases separately:

Case 1 $U_{t-1}^{i,\hat{g}} = 0$.

Case 2 $U_{t-1}^{i,\hat{g}} = 1$.

Case 1 When $U_{t-1}^{i,\hat{g}} = 0$, we must have $\overline{X}_{t-1}^{i,\hat{g}} = 0$ by (B.110). Then $\overline{X}_{t-1}^{j,\hat{g}} \in \{0, 1\}$ for the following reason. When $U_{t-1}^{i,\hat{g}} = U_{t-1}^{j,\hat{g}} = 0$, the sizes of both queues are between the lower bound and the threshold. That is

$$\overline{LB}_{t-1}^{\hat{g}} \leq \overline{X}_{t-1}^{i,\hat{g}} \leq \lceil TH_t \rceil - 1, \tag{B.111}$$

$$\overline{LB}_{t-1}^{\hat{g}} \leq \overline{X}_{t-1}^{j,\hat{g}} \leq \lceil TH_t \rceil - 1. \tag{B.112}$$

Combining (B.111), (B.112) with $\overline{X}_{t-1}^{i,\hat{g}} = 0$ we obtain

$$\begin{aligned} \overline{X}_{t-1}^{j,\hat{g}} &= \left| \overline{X}_{t-1}^{j,\hat{g}} - \overline{X}_{t-1}^{i,\hat{g}} \right| \\ &\leq \lceil TH_t \rceil - 1 - \overline{LB}_{t-1}^{\hat{g}} \\ &\leq \frac{1}{2} \left(\overline{UB}_{t-1}^{\hat{g}} - \overline{LB}_{t-1}^{\hat{g}} \right) \leq 1.5, \end{aligned} \tag{B.113}$$

where the last inequality in (B.113) is true because of (3.31) in Lemma III.4, (3.89), and

$$\overline{UB}_{t-1}^{\hat{g}} - \overline{LB}_{t-1}^{\hat{g}} \leq UB_t^{\hat{g}} + 1 - LB_t^{\hat{g}} + 1 \leq 3.$$

Therefore, $\overline{X}_{t-1}^{j,\hat{g}} \leq 1$ because $\overline{X}_{t-1}^{j,\hat{g}}$ takes integer values.

Case 2 When $U_{t-1}^{i,\hat{g}} = 1$, we must have $\overline{X}_{t-1}^{i,\hat{g}} = 1$ by (B.110). This implies that the threshold is not more than 1, and the only possible value of $\overline{X}_{t-1}^{j,\hat{g}}$ less than the threshold is 0.

As a consequence of the above analysis for the cases 1 and 2, $\{X_t^{i,\hat{g}} = 0\}$ implies

$$S_t = \overline{X}_{t-1}^{i,\hat{g}} + \overline{X}_{t-1}^{j,\hat{g}} \leq 1. \tag{B.114}$$

Thus, for $i = 1, 2$,

$$\left\{X_t^{i,\hat{g}} = 0\right\} = \left\{X_t^{i,\hat{g}} = 0, S_t \leq 1\right\}. \tag{B.115}$$

Then,

$$\begin{aligned}
&D_t^1 1_{\left\{X_t^{1,\hat{g}}=0\right\}} + D_t^2 1_{\left\{X_t^{2,\hat{g}}=0\right\}} \\
=&D_t^1 1_{\left\{X_t^{1,\hat{g}}=0, S_t \leq 1\right\}} + D_{t+1}^2 1_{\left\{X_t^{2,\hat{g}}=0, S_t \leq 1\right\}} \\
=&D_t^1 1_{\left\{X_t^{1,\hat{g}}=0, S_t=1\right\}} + D_t^2 1_{\left\{X_t^{1,\hat{g}}\neq 0, S_t=1\right\}} \\
&+ D_t^1 1_{\{S_t=0\}} + D_t^2 1_{\{S_t=0\}}.
\end{aligned} \tag{B.116}$$

Combining (B.109) and (B.116) we obtain

$$\begin{aligned}
&X_{t+1}^{1,\hat{g}} + X_{t+1}^{2,\hat{g}} \\
=&X_t^{1,\hat{g}} + X_t^{2,\hat{g}} - D_t^1 - D_t^2 + A_t^1 + A_t^2 \\
&+ D_t^1 1_{\left\{X_t^1=0, S_t=1\right\}} + D_t^2 1_{\left\{X_t^1\neq 0, S_t=1\right\}} \\
&+ D_t^1 1_{\{S_t=0\}} + D_t^2 1_{\{S_t=0\}} \\
=&S_{t+1},
\end{aligned} \tag{B.117}$$

where the last equality follows by the definition of S_{t+1}.

Therefore, at any time $t \geq T_0 + 1$ we have

$$X_t^{1,\hat{g}} + X_t^{2,\hat{g}} = S_t. \tag{B.118}$$

The proof of claim (3.90), and consequently, the proof of Theorem III.14 is complete.

□

Proof of Lemma III.4. This lemma follows from the system dynamics and the fact that for $i = 1, 2$ and any $k = 0, 1, \ldots, K$

$$\{U_t^{i,\mathrm{DR}_M} = k\} = \{\gamma_t(k) \leq \overline{X}_t^{i,\mathrm{DR}_M} \leq \gamma_t(k+1) - 1\}. \tag{B.119}$$

□

Proof of Lemma III.25. Assuming all variables in the proof are generated from the policy DR_M. From (3.98) and (3.101) we have

$$\begin{aligned} UB_{t+1}^i =& \gamma_t(U_t^i + 1) - U_t^i + U_t^j - 1 \\ =& \overline{LB}_t + \left\lfloor (2U_t^i + M + 1) \max\left(1, \frac{\overline{UB}_t - \overline{LB}_t + 1}{2K + M + 1}\right) \right\rfloor - U_t^i + U_t^j - 1 \\ =& \overline{LB}_t + U_t^i + U_t^j + M + \left\lfloor (2U_t^i + M + 1) \left(\max\left(1, \frac{\overline{UB}_t - \overline{LB}_t + 1}{2K + M + 1}\right) - 1\right) \right\rfloor \\ =& \overline{LB}_t + U_t^i + U_t^j + M + \left\lfloor \frac{2U_t^1 + M + 1}{2K + M + 1} \left(\overline{UB}_t - \overline{LB}_t - 2K - M\right)^+ \right\rfloor. \end{aligned} \tag{B.120}$$

Similarly, when $U_t^i \neq 0$ we have

$$\begin{aligned} & LB_{t+1}^i \\ =& \overline{LB}_t + U_t^i + U_t^j + M - 1 + \left\lfloor \frac{2U_t^1 + M - 1}{2K + M + 1} \left(\overline{UB}_t - \overline{LB}_t - 2K - M\right)^+ \right\rfloor; \end{aligned} \tag{B.121}$$

when $U_t^i = 0$,

$$LB_{t+1}^i = \overline{LB}_t + U_t^i + U_t^j. \tag{B.122}$$

Let $U_t^1 \geq U_t^2$ without loss of generality. Then (B.120)-(B.122) implies that

$$UB_{t+1} = UB_{t+1}^1, LB_{t+1} = LB_{t+1}^2, \tag{B.123}$$

Therefore,

$$\begin{aligned} UB_{t+1} - LB_{t+1} \leq & UB_{t+1} - (\overline{LB}_t + U_t^i + U_t^j) \\ = & M + \left\lfloor \frac{2U_t^1 + M + 1}{2K + M + 1} \left(\overline{UB}_t - \overline{LB}_t - 2K - M\right)^+ \right\rfloor, \end{aligned} \tag{B.124}$$

where the above inequality follows from (B.121) and (B.122), and the equality follows from (B.120).

Note that from (3.31),

$$\overline{UB}_t - \overline{LB}_t = UB_t + K - (LB_t - K)^+ \leq UB_t - LB_t + 2K \tag{B.125}$$

Putting (B.125) into (B.124) we obtain

$$\begin{aligned} UB_{t+1} - LB_{t+1} - M \leq & \left\lfloor \frac{2U_t^1 + M + 1}{2K + M + 1} \left(\overline{UB}_t - \overline{LB}_t - 2K - M\right)^+ \right\rfloor \\ \leq & \left\lfloor \frac{2U_t^1 + M + 1}{2K + M + 1} \left(UB_t - LB_t - M\right)^+ \right\rfloor \\ \leq & \left(UB_t - LB_t - M\right)^+ \end{aligned} \tag{B.126}$$

where the last inequality holds because $U_t^1 \leq K$.

□

Proof of Theorem III.27. Assuming all variables in the proof are generated from the policy DR_M.

Let $Y_t = (X_t^i, \overline{X}_t^i, UB_t^i, LB_t^i, \overline{UB}_t^i, \overline{LB}_t^i, i = 1, 2)$ be the variable including all states and bounds for the two queues at time t. It is clear that $\{Y_t, t \in \mathcal{Z}_+\}$ is a Markov

chain. Define a Lyapunov function $h(y_t)$ for $\{Y_t\}$ to be

$$h(y_t) = \overline{X}_t^1 + \overline{X}_t^2 + (2K+1)(UB_t - LB_t) \tag{B.127}$$

We claim that

$$\mathbb{E}\left[h(Y_{t+1}) - Y_t | Y_t = y_t\right] \leq -\epsilon \tag{B.128}$$

for some $\epsilon > 0$ and for all $y_t \notin C$, where $C = \{y_t : ub_t - lb_t > M\}$. Suppose (B.128) is true, then from Theorem 11.3.4 in [93] we obtain

$$\mathbb{E}[\tau_M] = \mathbb{E}[\mathbb{E}[\tau_M | Y_0]] \leq \mathbb{E}\left[\frac{1}{\epsilon} h(Y_0) | Y_0 \notin C\right] \leq \frac{1}{\epsilon}(2(UB_0 + K) + (2K+1)UB_0) < \infty \tag{B.129}$$

To prove (B.128), we consider two cases: $\max(U_t^1, U_t^2) = K$ and $\max(U_t^1, U_t^2) \leq K-1$. First consider the case when $\max(U_t^1, U_t^2) = K$.

Let $U_t^1 = K$ without loss of generality. Then

$$X_{t+1}^2 = \overline{X}_t^2 - U_t^2 + U_t^1 \geq U_t^1 = K. \tag{B.130}$$

Furthermore, we have $\overline{X}_t^1 \geq \gamma_t(K)$ because of $U_t^1 = K$. Consequently,

$$\begin{aligned} X_{t+1}^1 =& \overline{X}_t^1 - U_t^1 + U_t^2 \\ \geq& \gamma_t(K) - K \\ =& \overline{LB}_t + \left\lfloor (2K + M - 1) \max\left(1, \frac{\overline{UB}_t - \overline{LB}_t + 1}{2K + M + 1}\right) \right\rfloor - K \\ \geq& (2K + M - 1) - K \geq K. \end{aligned} \tag{B.131}$$

Since $X_{t+1}^i \geq K$, $(X_{t+1}^i - D_t^i)^+ = X_{t+1}^i - D_t^i$ for $i = 1, 2$. Therefore,

$$\begin{aligned}
&(\overline{X}_{t+1}^1 + \overline{X}_{t+1}^2) - (\overline{X}_t^2 + \overline{X}_t^2) \\
=&X_{t+1}^1 - D_t^1 + A_t^1 + X_{t+1}^1 - D_t^2 + A_t^2 - \overline{X}_t^1 - \overline{X}_t^2 \\
=&A_t^1 + A_t^2 - D_t^1 - D_t^2. \qquad \text{(B.132)}
\end{aligned}$$

where the last equality follows by (3.1) and (3.2).

For the second part of the function $h(\cdot)$, Lemma III.25 ensures, for $Y_t \notin C$,

$$\begin{aligned}
&(UB_{t+1} - LB_{t+1}) - (UB_t - LB_t) \\
=&(UB_{t+1} - LB_{t+1} - M) - (UB_t - LB_t - M) \\
\leq&(UB_t - LB_t - M)^+ - (UB_t - LB_t - M) = 0. \qquad \text{(B.133)}
\end{aligned}$$

Consequently, under the case when $y_t \notin C$ and $\max(U_t^1, U_t^2) = K$,

$$\begin{aligned}
&\mathbb{E}\left[h(Y_{t+1}) - Y_t | Y_t = y_t\right] \\
\leq&\mathbb{E}\left[A_t^1 + A_t^2 - D_t^1 - D_t^2\right] \\
=&-(\mu^1 + \mu^2 - \lambda^1 - \lambda^2). \qquad \text{(B.134)}
\end{aligned}$$

Now consider the second case when $\max(U_t^1, U_t^2) \leq K - 1$. From Lemma III.25 we have

$$\begin{aligned}
UB_{t+1} - LB_{t+1} - M &\leq \left\lfloor \frac{2\max(U_t^1, U_t^2) + M + 1}{2K + M + 1} (UB_t - LB_t - M)^+ \right\rfloor \\
&\leq \left\lfloor \frac{2K + M - 1}{2K + M + 1} (UB_t - LB_t - M)^+ \right\rfloor \qquad \text{(B.135)}
\end{aligned}$$

Using (B.135), for any $Y_t \notin C$, we get

$$\begin{aligned}
&(UB_{t+1} - LB_{t+1}) - (UB_t - LB_t) \\
\leq & M + \left\lfloor \frac{2K+M-1}{2K+M+1} (UB_t - LB_t - M)^+ \right\rfloor - (UB_t - LB_t) \\
= & \left\lfloor \frac{2K+M-1}{2K+M+1} (UB_t - LB_t - M) \right\rfloor - (UB_t - LB_t - M) \\
= & \left\lfloor \frac{-2}{2K+M+1} (UB_t - LB_t - M) \right\rfloor \\
\leq & -1
\end{aligned} \tag{B.136}$$

On the other hand, from system dynamics we have

$$\begin{aligned}
&(\overline{X}^1_{t+1} + \overline{X}^2_{t+1}) - (\overline{X}^1_t + \overline{X}^2_t) \\
= & (X^1_{t+1} - D^1_t)^+ + A^1_t + (X^2_{t+1} - D^2_t)^+ + A^2_t - \overline{X}^1_t - \overline{X}^2_t \\
\leq & X^1_{t+1} + A^1_t + X^2_{t+1} + A^2_t - \overline{X}^1_t - \overline{X}^2_t \\
= & A^1_t + A^2_t
\end{aligned} \tag{B.137}$$

where the last equality follows by (3.1) and (3.2).

Therefore, under the case when $y_t \notin C$ and $\max(U^1_t, U^2_t) \leq K-1$,

$$\begin{aligned}
&\mathbb{E}\left[h(Y_{t+1}) - Y_t | Y_t = y_t\right] \\
\leq & \mathbb{E}\left[A^1_t + A^2_t + (2K+1)(-1)\right] \\
= & (\lambda^1 + \lambda^2 - 2K) - 1 \leq -1
\end{aligned} \tag{B.138}$$

From the above analysis, (B.128) follows from (B.134) and (B.138) with

$$\epsilon = \min(1, \mu^1 + \mu^2 - \lambda^1 - \lambda^2). \tag{B.139}$$

□

APPENDIX C

Appendix for Multiple Access Communication

Proof of Lemma IV.1. The lemma is proved by induction. Equation (4.7) is true at $t = 0$ because all queues are initially empty. Suppose (4.7) is true at $t = k$. At time $t = k + 1$ we have

$$\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_{k+1}|f_{0:k}) = \frac{\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_{k+1}, f_k|f_{0:k-1})}{\sum_{q'_{k+1}} \mathbb{P}^{\lambda,\mathrm{CIMA}}(q'_{k+1}, f_k|f_{0:k-1})}. \tag{C.1}$$

Let $v = v(b_t^{\mathrm{CIMA}})$. Consider the numerator in (C.1). There are two cases: $f_k = 1$ and $f_k = 0$.

When $f_k = 1$, we have

$$
\begin{aligned}
&\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_{k+1}, F_k = 1 | f_{0:k-1}) \\
=&\sum_{q_k} \mathbb{P}^{\lambda,\mathrm{CIMA}}(q_{k+1}, q_k, F_k = 1 | f_{0:k-1}) \\
=&\sum_{q_k} \mathbb{P}^{\lambda,\mathrm{CIMA}}(q_{k+1}, q_k, Q_k^v > 0 | f_{0:k-1}) && \text{(C.2)} \\
=&\sum_{q_k, q_k^v > 0} \left[\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k | f_{0:k-1}) \mathbb{P}^{\lambda}(A_k^v = q_{k+1}^v - q_k^v + 1) \prod_{n \neq v} \mathbb{P}^{\lambda}(A_k^n = q_{k+1}^n - q_k^n) \right] && \text{(C.3)} \\
=&\sum_{q_k, q_k^v > 0} \left[\prod_{n=1}^{N} \mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^n | f_{0:k-1}) \mathbb{P}^{\lambda}(A_k^v = q_{k+1}^v - q_k^v + 1) \prod_{n \neq v} \mathbb{P}^{\lambda}(A_k^n = q_{k+1}^n - q_k^n) \right] \\
& && \text{(C.4)} \\
=&\prod_{n \neq v} \left[\sum_{q_k^n} \mathbb{P}^{\lambda}(A_k^n = q_{k+1}^n - q_k^n) \mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^n | f_{0:k-1}) \right] \\
&\sum_{q_k^v > 0} \mathbb{P}^{\lambda}(A_k^v = q_{k+1}^v - q_k^v + 1) \mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^v | f_{0:k-1}). && \text{(C.5)}
\end{aligned}
$$

Equation (C.2) holds because $F_k = 1$ if and only if $Q_k^{v(B_t^{\mathrm{CIMA}})} > 0$. Equation (C.3) is true because of the system dynamics (4.1) and the fact that $A_k^n, n = 1, 2, \ldots, N$ are mutually independent and independent of all variables before k. Equation (C.4) follows from the induction hypothesis for (4.7). Equation (C.5) is true because each term in (C.4) depends only on each q_k^n for $n = 1, 2, \ldots, N$.

Using (C.5), the denominator in (C.1) becomes

$$\begin{aligned}
&\sum_{q'_{k+1}}\left\{\prod_{n\neq v}\left[\sum_{q_k^n}\mathbb{P}^{\lambda}(A_k^n=q'^n_{k+1}-q_k^n)\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^n|f_{0:k-1})\right]\right.\\
&\qquad\left.\sum_{q_k^v>0}\mathbb{P}^{\lambda}(A_k^v=q'^v_{k+1}-q_k^v+1)\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^v|f_{0:k-1})\right\}\\
=&\sum_{q_k^v>0}\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^v|f_{0:k-1})\\
=&\mathbb{P}^{\lambda,\mathrm{CIMA}}(Q_k^v>0|f_{0:k-1});
\end{aligned}\tag{C.6}$$

equation (C.6) is true because all possible values of $q_k^n, n\neq v$ are summed out. Substituting (C.5) and (C.6) back into (C.1) we obtain for $f_k=1$

$$\begin{aligned}
&\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_{k+1}|f_{0:k})\\
=&\prod_{n\neq v}\left[\sum_{q_k^n}\mathbb{P}^{\lambda}(A_k^n=q_{k+1}^n-q_k^n)\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^n|f_{0:k-1})\right]\\
&\left[\sum_{q_k^v>0}\mathbb{P}^{\lambda}(A_k^v=q_{k+1}^v-q_k^v+1)\frac{\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^v|f_{0:k-1})}{\mathbb{P}^{\lambda,\mathrm{CIMA}}(Q_k^v>0|f_{0:k-1})}\right]\\
=:&\prod_{n=1}^{N}\phi^n(q_{k+1}^n),
\end{aligned}\tag{C.7}$$

where, for $n\neq v$,

$$\begin{aligned}
&\phi^n(q_{k+1}^n)\\
:=&\sum_{q_k^n}\mathbb{P}^{\lambda}(A_k^n=q_{k+1}^n-q_k^n)\mathbb{P}^{\lambda,\mathrm{CIMA}}(q_k^n|f_{0:k-1})\\
=&\mathbb{P}^{\lambda}(A_k^n=1)\mathbb{P}^{\lambda,\mathrm{CIMA}}(Q_k^n=q_{k+1}^n-1|f_{0:k-1})+\mathbb{P}^{\lambda}(A_k^n=0)\mathbb{P}^{\lambda,\mathrm{CIMA}}(Q_k^n=q_{k+1}^n|f_{0:k-1})\\
=&\lambda^n\mathbb{P}^{\lambda,\mathrm{CIMA}}(Q_k^n=q_{k+1}^n-1|f_{0:k-1})+(1-\lambda^n)\mathbb{P}^{\lambda,\mathrm{CIMA}}(Q_k^n=q_{k+1}^n|f_{0:k-1}),
\end{aligned}\tag{C.8}$$

and for $n = v$,

$$\begin{aligned}
&\phi^v(q^v_{k+1}) \\
:=&\sum_{q^v_k>0} \mathbb{P}^\lambda(A^v_k = q^v_{k+1} - q^v_k + 1)\frac{\mathbb{P}^{\lambda,\text{CIMA}}(q^v_k|f_{0:k-1})}{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k > 0|f_{0:k-1})} \\
=&\mathbb{P}^\lambda(A^v_k = 1)\frac{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k = q^v_{k+1}|f_{0:k-1})\mathbf{1}_{\{q^n_{t+1}>0\}}}{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k > 0|f_{0:k-1})} + \mathbb{P}^\lambda(A^v_k = 0)\frac{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k = q^v_{k+1}+1|f_{0:k-1})}{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k > 0|f_{0:k-1})} \\
=&\lambda^v\frac{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k = q^v_{k+1}|f_{0:k-1})\mathbf{1}_{\{q^v_{k+1}>0\}}}{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k > 0|f_{0:k-1})} + (1-\lambda^v)\frac{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k = q^v_{k+1}+1|f_{0:k-1})}{\mathbb{P}^{\lambda,\text{CIMA}}(Q^v_k > 0|f_{0:k-1})}.
\end{aligned} \tag{C.9}$$

Equations (C.8) and (C.9) follow from (C.7) and the fact that A^n_k takes values in $\{0, 1\}$ for all $n = 1, 2, \ldots, N$.

From (C.8) and (C.9) we conclude that $\phi^n(q^n_{k+1})$ is a probability mass function (PMF) for all n. This along with (C.7) establish that the marginal conditional PMF satisfies

$$\mathbb{P}^{\lambda,\text{CIMA}}(q^n_{k+1}|f_{0:k}) = \phi^n(q^n_{k+1}) \tag{C.10}$$

for all n, and this establish the induction step when $f_k = 1$.

When $f_k = 0$, by arguments similar to those in (C.2)-(C.10), we get

$$\mathbb{P}^{\lambda,\text{CIMA}}(q_{k+1}|f_{0:k}) = \prod_{n=1}^{N} \mathbb{P}^{\lambda,\text{CIMA}}(q^n_{k+1}|f_{0:k}), \tag{C.11}$$

where for $n \neq v$, by an argument similar to (C.8), we get

$$\begin{aligned}
&\mathbb{P}^{\lambda,\text{CIMA}}(q^n_{k+1}|f_{0:k}) \\
=&\lambda^n\mathbb{P}^{\lambda,\text{CIMA}}(Q^n_k = q^n_{k+1} - 1|f_{0:k-1}) + (1-\lambda^n)\mathbb{P}^{\lambda,\text{CIMA}}(Q^n_k = q^n_{k+1}|f_{0:k-1}),
\end{aligned} \tag{C.12}$$

and for $n = v$, by an argument similar to (C.9), we obtain

$$\mathbb{P}^{\lambda,\mathrm{CIMA}}(q^v_{k+1}|f_{0:k}) = \lambda^v \mathbf{1}_{\{q^v_{k+1}=1\}} + (1-\lambda^v)\mathbf{1}_{\{q^v_{k+1}=0\}}. \tag{C.13}$$

Therefore, the induction step is complete and (4.7) holds for all t. Furthermore, (4.8) is established by (C.8) and (C.12); (4.9) is established by (C.9), and (4.10) is established by (C.13).

□

Proof of Lemma IV.2. Equation (4.11) follows directly from (4.1), the queue length dynamics, and (4.6), the definition of the CIMA protocol.

For the common upper bounds, let $v = v(B^{\mathrm{CIMA}}_t)$, which is a function of $F_{0:t-1}$.

For $n \neq v$, we get $B^{n,\mathrm{CIMA}}_{t+1} = B^{n,\mathrm{CIMA}}_t + 1$ form (4.8) in Lemma IV.1.

For $n = v$ and $F_t = 1$, we obtain $B^{v,\mathrm{CIMA}}_{t+1} = B^{v,\mathrm{CIMA}}_t$ from (4.9) in Lemma IV.1.

For $n = v$ and $F_t = 0$, (4.10) in Lemma IV.1 gives $B^{v,\mathrm{CIMA}}_{t+1} = 1$, and the proof of the lemma is complete.

□

Proof of Lemma IV.4. From (4.11) and (4.12) in Lemma IV.2 we know that Q^{CIMA}_{t+1} and B^{CIMA}_{t+1} are functions of $Q^{\mathrm{CIMA}}_t, B^{\mathrm{CIMA}}_t$, A^n_t and F_t. From (4.6), the definition of the CIMA protocol, we know that

$$F_t = U^{v(B^{\mathrm{CIMA}}_t)}_t = \mathbf{1}_{\left\{Q^{v(B^{\mathrm{CIMA}}_t),\mathrm{CIMA}}_t > 0\right\}}.$$

Therefore, F_t is a function of Q^{CIMA}_t and B^{CIMA}_t. Consequently, Y^{CIMA}_{t+1} is a function

of $Q_t^{\text{CIMA}}, B_t^{\text{CIMA}}$ and A_t^n. Let $f(Y_t^{\text{CIMA}}, A_t^n) := Y_{t+1}^{\text{CIMA}}$, we have

$$\begin{aligned}
&\mathbb{P}(Y_{t+1}^{\text{CIMA}} = y_{t+1} | Y_k^{\text{CIMA}} = y_k, k \le t) \\
=&\mathbb{P}(f(Y_t^{\text{CIMA}}, A_t^n) = y_{t+1} | Y_k^{\text{CIMA}} = y_k, k \le t) \\
=&\mathbb{P}(f(y_t, A_t^n) = y_{t+1} | Y_k^{\text{CIMA}} = y_k, k \le t) \\
\overset{(*)}{=}&\mathbb{P}(f(y_t, A_t^n) = y_{t+1} | Y_t^{\text{CIMA}} = y_t) \\
=&\mathbb{P}(Y_t^{\text{CIMA}} = y_{t+1} | Y_t^{\text{CIMA}} = y_t),
\end{aligned}$$

where (*) is true because A_t^n is independent of $Q_t^{\text{CIMA}}, B_t^{\text{CIMA}}$ and all random variables before time slot t.

Therefore, $\{Y_t^{\text{CIMA}}, t = 0, 1, \dots\}$ is a Markov chain.

□

Detailed derivation of (4.14) *in the proof of Theorem IV.3.* From the definition of the Lyapunov function $h(\cdot)$ (cf (4.13)) we have

$$\begin{aligned}
&\mathbb{E}\left[h(Y_{t+1}^{\text{CIMA}}) - h(Y_t^{\text{CIMA}}) | Y_t^{\text{CIMA}} = y\right] \\
=&\mathbb{E}\left[\sum_{n=1}^{N} (Q_{t+1}^{n,\text{CIMA}} - Q_t^{n,\text{CIMA}}) | Y_t^{\text{CIMA}} = y\right] + \alpha \mathbb{E}\left[\sum_{n=1}^{N} (B_{t+1}^{n,\text{CIMA}} - B_t^{n,\text{CIMA}}) | Y_t^{\text{CIMA}} = y\right].
\end{aligned} \tag{C.14}$$

For $n \neq v$, from (4.11) and (4.12) in Lemma IV.2, we get

$$\begin{aligned}
&\mathbb{E}\left[(Q_{t+1}^{n,\text{CIMA}} - Q_t^{n,\text{CIMA}}) | Y_t^{\text{CIMA}} = y\right] + \alpha \mathbb{E}\left[(B_{t+1}^{n,\text{CIMA}} - B_t^{n,\text{CIMA}}) | Y_t^{\text{CIMA}} = y\right] \\
=&\mathbb{E}\left[A_t^n | Y_t^{\text{CIMA}} = y\right] + \alpha \mathbb{E}\left[1 | Y_t^{\text{CIMA}} = y\right] \\
=&\lambda^n + \alpha,
\end{aligned} \tag{C.15}$$

where the last equality in (C.15) holds because A_t^n is independent of Y_t^{CIMA}.

For $n = v$, from (4.11) and (4.12) in Lemma IV.2, we obtain

$$\begin{aligned}
&\mathbb{E}\left[(Q_{t+1}^{v,\mathrm{CIMA}} - Q_t^{v,\mathrm{CIMA}})|Y_t^{\mathrm{CIMA}} = y\right] + \alpha\mathbb{E}\left[(B_{t+1}^{v,\mathrm{CIMA}} - B_t^{v,\mathrm{CIMA}})|Y_t^{\mathrm{CIMA}} = y\right] \\
=&\mathbb{E}\left[A_t^v - 1 + 1_{\{Q_t^v=0\}}|Y_t^{\mathrm{CIMA}} = y\right] + \alpha\mathbb{E}\left[(1 - B_t^v)1_{\{Q^v=0\}}|Y_t^{\mathrm{CIMA}} = y\right] \\
=&\mathbb{E}\left[A_t^v - 1 + 1_{\{q^v=0\}}|Y_t^{\mathrm{CIMA}} = y\right] + \alpha\mathbb{E}\left[(1 - b^v)1_{\{q^v=0\}}|Y_t^{\mathrm{CIMA}} = y\right] \\
=&\lambda^v - 1 + (1 + \alpha(1 - b^v))1_{\{q^v=0\}}, \qquad \text{(C.16)}
\end{aligned}$$

where the last equality in (C.16) follows from the fact that A_t^v is also independent of Y_t^{CIMA}.

substituting (C.15) and (C.16) back into (C.14) we get

$$\begin{aligned}
&\mathbb{E}\left[h(Y_{t+1}^{\mathrm{CIMA}}) - h(Y_t^{\mathrm{CIMA}})|Y_t^{\mathrm{CIMA}} = y\right] \\
=&\sum_{n\neq v} \lambda^n + \alpha(N-1) + \lambda^v - 1 + (1 + \alpha(1 - b^v))1_{\{q^v=0\}} \\
\overset{(a)}{=}& -\epsilon + \alpha(N-1) + (1 + \alpha(1 - b^v))1_{\{q^v=0\}} \\
\overset{(b)}{=}& -\epsilon/2 + (1 + \alpha(1 - b^v))1_{\{q^v=0\}} \\
\leq& -\epsilon/2 \quad \text{if } b^v \geq \frac{1}{\alpha} + 1, \qquad \text{(C.17)}
\end{aligned}$$

where (a) in (C.17) is true because $\sum_{n=1}^{N} \lambda^n = 1 - \epsilon$, and (b) in (C.17) is true because $\alpha = \frac{\epsilon}{2(N-1)}$. Consequently, inequality (4.14) in the proof of Theorem IV.3 is established. □

Proof of Lemma IV.7. The lemma holds if there is no unsuccessful transmission from time t to $t + q + N - 1$. Otherwise, suppose the first unsuccessful transmission is from user n at time $t_1, t \leq t_1 \leq t + q + N - 1$. Since $v(B_{t_1}^{\mathrm{CIMA}}) = n$ and no packet is transmitted at time t_1, every user will update the upper bound $B_{t_1+1}^{n,\mathrm{CIMA}} = 1$ for user

n. From the evolution of the upper bounds we have

$$B_\tau^{n,\mathrm{CIMA}} = \tau - t_1 \tag{C.18}$$

for any time τ if user n is not selected again by CIMA before time τ.

There are two possibilities: (1) user n is not selected by CIMA again before time $t+q+N$; (2) user n is selected by CIMA again at time t_2 where $t_1+1 \le t_2 \le t+q+N-1$.

First consider the case when user n is not selected by CIMA again before time $t+q+N$. Then (C.18) holds for any time τ from t_1+1 to $t+q+N-1$. From the specification of CIMA, if any other user m has an unsuccessful transmission at time t', $t_1+1 \le t' \le t+q+N-1$, for any subsequent time $\tau \ge t'+1$ we will have

$$B_\tau^{m,\mathrm{CIMA}} \le \tau - t' < \tau - t_1 = B_\tau^{n,\mathrm{CIMA}}. \tag{C.19}$$

Therefore, user m will not be selected by CIMA again from time $t'+1$ to $t+q+N-1$. Consequently, any user $m \neq n$ can have at most one unsuccessful transmission from time t to $t+q+N-1$. Since any of the N users can have at most one unsuccessful transmission from time t to $t+q+N-1$, the number of successful transmissions during this time period is at least $(t+q+N-1)-t+1-N = q$.

Next consider the case when user n is selected by CIMA again at time t_2 where $t_1+1 \le t_2 \le t+q+N-1$. From the specification of CIMA, we must have $B_{t_2}^{n,\mathrm{CIMA}} = \max_m B_{t_2}^{m,\mathrm{CIMA}}$ for user n to transmit at time t_2. Therefore, letting $\tau = t_2$ in (C.18) we get

$$t_2 - t_1 = B_{t_2}^{n,\mathrm{CIMA}} \ge B_{t_2}^{m,\mathrm{CIMA}} \tag{C.20}$$

for all $m \neq n$. Let $S^m, m \neq n$ be the number of successful transmissions for user m

between time t_1 and t_2. We prove in the following that $S^m \geq Q_{t_1}^{m,\text{CIMA}}$.

If user m has an unsuccessful transmission between t_1 and t_2, then the queue at user m is empty at the time of the unsuccessful transmission. Therefore, $S^m \geq Q_{t_1}^{m,\text{CIMA}}$ because all the $Q_{t_1}^{m,\text{CIMA}}$ packets queued at time t_1 are successfully transmitted by user m between time t_1 and t_2.

If user m transmits successfully in every time slot selected by CIMA, from (C.20) and the evolution of the upper bounds we obtain

$$t_2 - t_1 \geq B_{t_2}^{m,\text{CIMA}} = B_{t_1}^{m,\text{CIMA}} + t_2 - t_1 - S^m. \tag{C.21}$$

Since $B_{t_1}^{m,\text{CIMA}} \geq Q_{t_1}^{m,\text{CIMA}}$, (C.21) implies

$$S^m \geq B_{t_1}^{m,\text{CIMA}} \geq Q_{t_1}^{m,\text{CIMA}}. \tag{C.22}$$

Consequently, for every user $m \neq n$

$$S^m \geq Q_{t_1}^{m,\text{CIMA}}. \tag{C.23}$$

Note that the total number of successful transmissions between t_1 and t_2 is $\sum_{m\neq n} S^m$; therefore,

$$\sum_{\tau=t_1+1}^{t_2-1} \bar{U}_\tau = \sum_{m\neq n} S^m \geq \sum_{m\neq n} Q_{t_1}^{m,\text{CIMA}} = Q_{t_1}^{tot,\text{CIMA}} \tag{C.24}$$

where the last equation in (C.24) holds because $Q_{t_1}^{n,\text{CIMA}} = 0$.

From the dynamics of queues we get

$$Q_{t_1}^{tot,\text{CIMA}} = Q_t^{tot,\text{CIMA}} + \sum_{\tau=t}^{t_1-1} \left(\sum_{n=1}^{N} A_\tau^n - \bar{U}_\tau \right) \geq q - \sum_{\tau=t}^{t_1-1} \bar{U}_\tau \tag{C.25}$$

Combining (C.24) and (C.25), the total number of successful transmissions from time

t to $t+q+N-1$ in the second case is at least

$$\begin{aligned}\sum_{\tau=t}^{t+q+N-1}\bar{U}_\tau \geq & \sum_{\tau=t}^{t_1-1}\bar{U}_\tau + \sum_{\tau=t_1+1}^{t_2-1}\bar{U}_\tau \\ \geq & q - Q_{t_1}^{tot,\mathrm{CIMA}} + Q_{t_1}^{tot,\mathrm{CIMA}} = q. \end{aligned} \tag{C.26}$$

□

Detailed derivation of (4.19) *and* (4.21) *in the proof of Theorem IV.6. Detailed derivation of* (4.19)*:* From (4.18) and the dynamics of queues we obtain

$$\begin{aligned} Q_{T_k}^{tot,\mathrm{CIMA}} =& Q_{T_{k-1}}^{tot,\mathrm{CIMA}} + \sum_{t=T_{k-1}}^{T_k-1}\left(\sum_{n=1}^{N}A_t^n - \bar{U}_t^{\mathrm{CIMA}}\right) \\ =& Q_{T_{k-1}}^{tot,\mathrm{CIMA}} - \sum_{t=T_{k-1}}^{T_k-1}\bar{U}_t^{\mathrm{CIMA}} + \sum_{t=T_{k-1}}^{T_k-1}\sum_{n=1}^{N}A_t^n \\ =& \sum_{t=T_{k-1}}^{T_k-1}\sum_{n=1}^{N}A_t^n \qquad (C.27) \\ \leq & \sum_{t=T_{k-1}}^{T_{k-1}+Q_{T_{k-1}}^{tot,\mathrm{CIMA}}+N-1}\sum_{n=1}^{N}A_t^n. \qquad (C.28) \end{aligned}$$

Equation (C.27) follows from the definition of T_k. Inequality (C.28) is true because of (4.18) and the fact that A_t^n are all positive. Note that A_t^n is independent of T_{k-1} and $Q_{T_{k-1}}^{tot,\mathrm{CIMA}}$ for $t \geq T_{k-1}$. Therefore, taking the expectation on both sides of (C.28) we get

$$
\begin{aligned}
\mathbb{E}\left[Q_{T_k}^{tot,\mathrm{CIMA}}\right] \leq & \mathbb{E}\left[\sum_{t=T_{k-1}}^{T_{k-1}+Q_{T_{k-1}}^{tot,\mathrm{CIMA}}+N-1}\sum_{n=1}^{N} A_t^n\right] \\
= & \mathbb{E}\left[\mathbb{E}\left[\sum_{t=T_{k-1}}^{T_{k-1}+Q_{T_{k-1}}^{tot,\mathrm{CIMA}}+N-1}\sum_{n=1}^{N} A_t^n \middle| T_{k-1}, Q_{T_{k-1}}^{tot,\mathrm{CIMA}}\right]\right] \\
= & \mathbb{E}\left[\sum_{t=T_{k-1}}^{T_{k-1}+Q_{T_{k-1}}^{tot,\mathrm{CIMA}}+N-1}\sum_{n=1}^{N} \mathbb{E}\left[A_t^n | T_{k-1}, Q_{T_{k-1}}^{tot,\mathrm{CIMA}}\right]\right] \\
= & \mathbb{E}\left[\sum_{t=T_{k-1}}^{T_{k-1}+Q_{T_{k-1}}^{tot,\mathrm{CIMA}}+N-1}\sum_{n=1}^{N} \lambda^n\right] \\
= & \lambda^{tot}\left(\mathbb{E}\left[Q_{T_{k-1}}^{tot,\mathrm{CIMA}}\right]+N\right).
\end{aligned} \tag{C.29}
$$

Detailed derivation of (4.21)*:*

For any time $t = 0, 1, 2, \ldots$, suppose $T_{k-1} < t \leq T_k$ ($T_0 := 0$). Using (4.19) and the dynamics of queues we get

$$
\begin{aligned}
\mathbb{E}\left[Q_t^{tot,\mathrm{CIMA}}\right] = & \mathbb{E}\left[Q_{T_{k-1}}^{tot,\mathrm{CIMA}} + \sum_{\tau=T_{k-1}}^{t-1}\left(\sum_{n=1}^{N} A_\tau^n - \bar{U}_\tau^{\mathrm{CIMA}}\right)\right] \\
\leq & \mathbb{E}\left[Q_{T_{k-1}}^{tot,\mathrm{CIMA}} + \sum_{\tau=T_{k-1}}^{t-1}\left(\sum_{n=1}^{N} A_\tau^n\right)\right] \\
\leq & \mathbb{E}\left[Q_{T_{k-1}}^{tot,\mathrm{CIMA}} + \sum_{\tau=T_{k-1}}^{T_k-1}\left(\sum_{n=1}^{N} A_\tau^n\right)\right] \\
\overset{(a)}{=} & \mathbb{E}\left[Q_{T_{k-1}}^{tot,\mathrm{CIMA}} + Q_{T_k}^{tot,\mathrm{CIMA}}\right] \\
\leq & 2\frac{\lambda^{tot} N}{1-\lambda^{tot}};
\end{aligned} \tag{C.30}
$$

(a) in (C.30) follows from (C.27) and the last inequality in (C.30) follows from (4.20).

□

APPENDIX D

Appendix for Dynamic Stochastic Games with Asymmetric Information

Proof of Lemma V.8. The lemma is proved by induction. Since the initial states are independent, (5.19) holds at $t = 1$. Suppose the lemma is true at time t. Given any $h^c_{t+1} = \{c_{1:t+1}, y_{1:t}, a_{1:t}\}$ at $t+1$, we have from Bayes' rule

$$\begin{aligned}
&\hat{\pi}_{t+1}(x_{t+1}) \\
=&\mathbb{P}_{(A_{1:t}=a_{1:t})}(x_{t+1}|y_{1:t}) \\
=&\frac{\mathbb{P}_{(A_{1:t}=a_{1:t})}(x_{t+1}, y_t|y_{1:t-1})}{\sum_{x'_{t+1}\in\mathcal{X}_{t+1}} \mathbb{P}_{(A_{1:t}=a_{1:t})}(x'_{t+1}, y_t|y_{1:t-1})}.
\end{aligned} \tag{D.1}$$

The numerator in (D.1) can be further expressed by

$$\begin{aligned}
&\mathbb{P}_{(A_{1:t}=a_{1:t})}(x_{t+1}, y_t|y_{1:t-1}) \\
=&\sum_{x_t\in\mathcal{X}_t} \mathbb{P}_{(A_{1:t}=a_{1:t})}(x_{t+1}, y_t, x_t|y_{1:t-1}) \\
\overset{(a)}{=}&\sum_{x_t\in\mathcal{X}_t}\prod_{n=1}^{N} p_t^n(x_{t+1}^n; x_t^n, a_t)q_t^n(y_t^n; x_t^n, a_t)\hat{\pi}_t^n(x_t^n) \\
=&\prod_{n=1}^{N}\sum_{x_t^n\in\mathcal{X}_t^n} p_t^n(x_{t+1}^n; x_t^n, a_t)q_t^n(y_t^n; x_t^n, a_t)\hat{\pi}_t^n(x_t^n).
\end{aligned} \tag{D.2}$$

Equation (a) in (D.2) follows from the system dynamics, the fact $A_t = a_t$ and the induction hypothesis for the lemma. Substituting (D.2) into both the numerator and denominator of (D.1) we get

$$\begin{aligned}
&\hat{\pi}_{t+1}(x_{t+1}) \\
=&\prod_{n=1}^{N}\frac{\sum_{x_t^n\in\mathcal{X}_t^n} p_t^n(x_{t+1}^n; x_t^n, a_t)q_t^n(y_t^n; x_t^n, a_t)\hat{\pi}_t^n(x_t^n)}{\sum_{x_t'^n\in\mathcal{X}_t^n} q_t^n(y_t^n; x_t'^n, a_t)\hat{\pi}_t^n(x_t'^n)} \\
=&\prod_{n=1}^{N}\hat{\psi}_t^n(y_t^n, a_t, \hat{\pi}_t)(x_{t+1}^n).
\end{aligned} \tag{D.3}$$

□

Proof of Lemma V.13. If λ is a CIB strategy profile and ψ is a CIB update rule consistent with λ, we define $g \in \mathcal{G}$ to be the strategy profile constructed by (5.27) from (λ, ψ).

We proceed to recursively define a belief system μ and maps $\{\mu_t^c, t \in \mathcal{T}\}$ that satisfy (5.31)-(5.32), and are such that μ is consistent with g. For that matter, we first define the signaling-free belief $\hat{\mu}_t^c : \mathcal{H}_t^c \mapsto \Delta(\mathcal{X}_{1:t})$ given $h_t^c = \{a_{1:t-1}, y_{1:t-1}\} \in \mathcal{H}_t^c$ such that for any $x_{1:t} \in \mathcal{X}_{1:t}$

$$\hat{\mu}_t^c(h_t^c)(x_{1:t}) := \mathbb{P}_{(A_{1:t-1}=a_{1:t-1})}(x_{1:t}|y_{1:t-1}). \tag{D.4}$$

At time $t = 1$ we define, for all $h_1^n = (x_1^n, h_1^c) \in \mathcal{H}_1^n, n \in \mathcal{N}$ and for all $x_1 \in \mathcal{X}_1$

$$\mu_1^c(h_1^c)(x_1) := \mathbb{P}(x_1) \tag{D.5}$$

$$\mu_1^n(h_1^n)(x_1) := \mathbf{1}_{\{x_1^n | h_1^n\}} \mathbb{P}(x_1^{-n}) \tag{D.6}$$

Then, (5.31) and (5.32) are satisfied at time 1, and g is consistent with μ before time 1. (basis of induction)

Suppose $\mu_t^c(h_t^c)(\cdot)$ and $\mu_t^n(h_t^n)(\cdot)$ are defined, (5.31) and (5.32) are satisfied at time t, and g is consistent with μ before time t (induction hypothesis).

We proceed to define $\mu_{t+1}^c(h_{t+1}^c)(\cdot)$, and $\mu_{t+1}^n(h_{t+1}^n)(\cdot)$, and prove that (5.31) and (5.32) are satisfied at time $t+1$, and g is consistent with μ before time $t+1$. We first define

$$\begin{aligned} &\eta_t^k(x_t^k, y_t^k, a_t, h_t^c) := \eta_t^k(x_t^k, y_t^k, a_t, (c_t, \gamma_{\psi,t}(h_t^c), \hat{\gamma}_t(h_t^c))) \\ =&q_t^k(y_t^k; x_t^k, a_t) \lambda_t^k(x_t^n, c_t, \gamma_{\psi,t}(h_t^c), \hat{\gamma}_t(h_t^c))(a_t^n). \end{aligned} \tag{D.7}$$

At time $t+1$, for any histories h_{t+1}^c and $h_{t+1}^n, n \in \mathcal{N}$, we define the beliefs

$$\mu_{t+1}^c(h_{t+1}^c)(x_{1:t+1}) := \prod_{k \in \mathcal{N}} \mu_{t+1}^c(h_{t+1}^c)(x_{1:t+1}^k), \tag{D.8}$$

$$\begin{aligned} &\mu_{t+1}^n(h_{t+1}^n)(x_{1:t+1}) \\ :=&\mathbf{1}_{\{x_{1:t+1}^n\}}(h_{t+1}^n) \prod_{k \neq n} \mu_{t+1}^c(h_{t+1}^c)(x_{1:t+1}^k), \end{aligned} \tag{D.9}$$

where for any $k \in \mathcal{N}$

$$\begin{aligned} &\mu_{t+1}^c(h_{t+1}^c)(x_{1:t+1}^k) \\ :=&\frac{p_t^k(x_{t+1}^k; x_t^k, a_t) \eta_t^k(x_t^k, y_t^k, a_t, h_t^c) \mu_t^c(h_t^c)(x_{1:t}^k)}{\sum_{x_t'^k \in \mathcal{X}_t^k} \eta_t^k(x_t'^k, y_t^k, a_t, h_t^c) \mu_t^c(h_t^c)(x_t'^k)} \end{aligned} \tag{D.10}$$

when the denominator of (D.10) is non-zero; when the denominator of (D.10) is zero, $\mu^c_{t+1}(h^c_{t+1})(x^k_{1:t+1})$ is defined by

$$\begin{aligned}
&\mu^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) \\
:=&\begin{cases}
0 \qquad \text{when } \hat{\mu}^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) = 0, \\
\frac{\gamma_{\psi,t+1}(h^c_{t+1})(x^k_{t+1})}{|\{x'^k_{1:t} \in \mathcal{X}^k_{1:t} : \hat{\mu}^c_{t+1}(h^c_{t+1})(x'^k_{1:t}, x^k_{t+1}) \neq 0\}|} \\
\qquad\qquad \text{when } \hat{\mu}^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) \neq 0.
\end{cases}
\end{aligned} \tag{D.11}$$

Then (5.31) at $t+1$ follows directly from the above construction. We proceed to prove (5.32) at $t+1$.

First consider the case when the denominator of (D.10) is zero. Then, for any $k \in \mathcal{N}$, we obtain, because of (D.11),

$$\begin{aligned}
&\mu^c_{t+1}(h^c_{t+1})(x^k_{t+1}) \\
=&\sum_{x^k_{1:t} \in \mathcal{X}^k_{1:t}} \mu^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) \\
=&\sum_{\substack{x^k_{1:t} \in \mathcal{X}^k_{1:t}: \\ \hat{\mu}^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) \neq 0}} \frac{\gamma_{\psi,t+1}(h^c_{t+1})(x^k_{t+1})}{|\{x'^k_{1:t} \in \mathcal{X}^k_{1:t} : \hat{\mu}^c_{t+1}(h^c_{t+1})(x'^k_{1:t}, x^k_{t+1}) \neq 0\}|} \\
=&\gamma_{\psi,t+1}(h^c_{t+1})(x^k_{t+1}).
\end{aligned} \tag{D.12}$$

When the denominator of (D.10) is non-zero, from (D.10) we get

$$
\begin{aligned}
&\mu^c_{t+1}(h^c_{t+1})(x^k_{t+1}) \\
=&\sum_{x^k_{1:t}\in\mathcal{X}^k_{1:t}} \mu^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) \\
=&\sum_{x^k_{1:t}\in\mathcal{X}^k_{1:t}} \frac{p^k_t(x^k_{t+1};x^k_t,a_t)\eta^k_t(x^k_t,y^k_t,a_t,h^c_t)\mu^c_t(h^c_t)(x^k_{1:t})}{\sum_{x'^k_t\in\mathcal{X}^k_t}\eta^k_t(x'^k_t,y^k_t,a_t,h^c_t)\mu^c_t(h^c_t)(x'^k_t)} \\
=&\sum_{x^k_t\in\mathcal{X}^k_t} \frac{p^k_t(x^k_{t+1};x^k_t,a_t)\eta^k_t(x^k_t,y^k_t,a_t,h^c_t)\mu^c_t(h^c_t)(x^k_t)}{\sum_{x'^k_t\in\mathcal{X}^k_t}\eta^k_t(x'^k_t,y^k_t,a_t,h^c_t)\mu^c_t(h^c_t)(x'^k_t)} \\
\overset{(a)}{=}&\sum_{x^k_t\in\mathcal{X}^k_t} \frac{p^k_t(x^k_{t+1};x^k_t,a_t)\eta^k_t(x^k_t,y^k_t,a_t,h^c_t)\gamma_{\psi,t}(h^c_t)(x^k_t)}{\sum_{x'^k_t\in\mathcal{X}^k_t}\eta^k_t(x'^k_t,y^k_t,a_t,h^c_t)\gamma_{\psi,t}(h^c_t)(x'^k_t)} \\
=&\gamma_{\psi,t+1}(h^c_{t+1})(x^k_{t+1}),
\end{aligned}
\tag{D.13}
$$

where (a) in (D.13) follows from the induction hypothesis for (5.31) at time t, and the last equality in (D.13) is true because of (5.28) (ψ is consistent with λ).

Therefore, (5.32) is true at time $t+1$ from (D.12) and (D.13).

To show consistency at time $t+1$, we need to show that Bayes' rule, given by (5.15), is satisfied when the denominator of (5.15) is non-zero, and (5.16) holds for any histories h^n_{t+1}.

We first note that, the construction (D.10), (D.11) and the definition of signaling-free belief $\hat{\mu}^c_{t+1}$ ensure that

$$
\mu^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) = 0 \text{ if } \hat{\mu}^c_{t+1}(h^c_{t+1})(x^k_{1:t+1}) = 0. \tag{D.14}
$$

Therefore, (5.16) follows from (D.14) since the signaling-free belief satisfies

$$
\begin{aligned}
&\hat{\mu}^n_{t+1}(h^n_{t+1})(x_{1:t+1}) \\
=&\mathbf{1}_{\{x^n_{1:t+1}\}}(h^n_{t+1})\prod_{k\neq n}\hat{\mu}^c_t(h^c_{t+1})(x^k_{1:t+1})
\end{aligned}
\tag{D.15}
$$

which follows by an argument similar to that of Lemma V.8.

Now consider (5.15) at $t+1$ when the denominator is non-zero. From (D.9)-(D.10) the left hand side of (5.15) equals to

$$\mu^n_{t+1}(h^n_{t+1})(x_{1:t+1}) = \mathbf{1}_{\{x^n_{1:t+1}\}}(h^n_{t+1}) \prod_{k\neq n} \frac{p^k_t(x^k_{t+1}; x^k_t, a_t)\eta^k_t(x^k_t, y^k_t, a_t, h^c_t)\mu^c_t(h^c_t)(x^k_{1:t})}{\sum_{x'^k_t \in \mathcal{X}^k_t} \eta^k_t(x'^k_t, y^k_t, a_t, h^c_t)\mu^c_t(h^c_t)(x'^k_t)}. \tag{D.16}$$

On the other hand, the numerator of the right hand side of (5.15) is equal to

$$\begin{aligned}
&\mathbb{P}^{g_t}_{\mu}(x_{1:t+1}, y_t, a_t | h^n_t, a^n_t) \\
=&\mu^n_t(h^n_t)(x_{1:t}) \prod_{k\in\mathcal{N}} p^k_t(x^k_{t+1}; x^k_t, a_t) q^k_t(y^k_t; x^k_t, a_t) \prod_{k\neq n} \lambda^k_t(x^k_t, c_t, \gamma_{\psi,t}(h^c_t), \hat{\gamma}_t(h^c_t))(a^k_t) \\
=&\mathbf{1}_{\{x^n_{1:t}\}}(h^n_t) \prod_{k\in\mathcal{N}} p^k_t(x^k_{t+1}; x^k_t, a_t) q^k_t(y^k_t; x^k_t, a_t) \\
&\prod_{k\neq n} \mu^c_t(h^c_t)(x^k_{1:t}) \prod_{k\neq n} \lambda^k_t(x^k_t, c_t, \gamma_{\psi,t}(h^c_t), \hat{\gamma}_t(h^c_t))(a^k_t) \\
=&\mathbf{1}_{\{x^n_{1:t}\}}(h^n_t) p^n_t(x^n_{t+1}; x^n_t, a_t) q^n_t(y^n_t; x^n_t, a_t) \\
&\prod_{k\neq n} p^k_t(x^k_{t+1}; x^k_t, a_t)\eta^k_t(x^k_t, y^k_t, a_t, h^c_t)\mu^c_t(h^c_t)(x^k_{1:t})
\end{aligned} \tag{D.17}$$

The first equality in (D.17) follows from (5.12) and (5.27). The second equality in (D.17) follows from the induction hypothesis for (5.31). The last equality in (D.17) follows from (D.7).

Substituting (D.17) back into both the numerator and the denominator in the right hand side of (5.15), we obtain (D.16). Therefore, (5.15) is satisfied for any history $h^n_{t+1} \in \mathcal{H}^n_{t+1}$ for any $n \in \mathcal{N}$ when the denominator of (5.15) is non-zero, hence, (g, μ) is consistence before time $t+1$. This completes the induction step and the proof of the lemma.

□

Proof of Lemma V.14. If agent n uses an arbitrary strategy g'^n, following the same construction (D.9)-(D.10) in Lemma V.13, we can obtain a belief system μ' from $g' := (g'^n, g^{-n})$ and ψ such that

$$\mu'^n_t(h^n_t)(x_{1:t}) = \mathbf{1}_{\{x^n_{1:t}\}}(h^n_t) \prod_{k \neq n} \mu'^c_t(h^c_t)(x^k_{1:t}). \tag{D.18}$$

Since $\mu'^c_t(h^c_t)(x^k_{1:t})$ defined by (D.10) and (D.11) depends only on the strategies $g'^{-n} = g^{-n}$ of all agents other than n, we have for all $h^c_t \in \mathcal{H}^c_t$

$$\mu'^c_t(h^c_t)(x^k_{1:t}) = \mu^c_t(h^c_t)(x^k_{1:t}). \tag{D.19}$$

Therefore, for any history $H^n_t \in \mathcal{H}^n_t$

$$\mu'^n_t(h^n_t)(x_{1:t}) = \mathbf{1}_{\{x^n_{1:t}\}}(h^n_t) \prod_{k \neq n} \mu^c_t(h^c_t)(x^k_{1:t}). \tag{D.20}$$

The same argument for the proof of consistency in Lemma V.13 shows that μ' is consistent with $g' = (g'^n, g^{-n})$. Therefore, when $\mathbb{P}^{g'^n, g^{-n}}(h^n_t) > 0$, from Bayes' rule we have

$$\begin{aligned} \mathbb{P}^{g'^n, g^{-n}}(x_{1:t}|h^n_t) =& \mathbb{P}^{g'^n, g^{-n}}_{\mu'}(x_{1:t}|h^n_t) = \mu'^n_t(h^n_t)(x_{1:t}) \\ =& \mathbf{1}_{\{x^n_{1:t}\}}(h^n_t) \prod_{k \neq n} \mu^c_t(h^c_t)(x^k_{1:t}). \end{aligned} \tag{D.21}$$

□

Proof of Lemma V.16. To simply the notation, we use Π_t to denote $\Pi^{\gamma_\psi}_t$ and $B_t = (C_t, \Pi_t, \hat{\Pi}_t)$.

Let $(g, \mu) = f(\lambda, \psi)$ as in Lemma V.13. Suppose every agent $k \neq n$ uses the strategy g^k along with the belief system μ.

Below, we show that agent n's best response problem (5.14) is a Markov Decision

Process (MDP) with state process $\{(X_t^n, B_t), t \in \mathcal{T}\}$ and action process $\{A_t^n, t \in \mathcal{T}\}$.

Since the strategies g^{-n} of all other agents are fixed, when agent n selects an action $a_t^n \in \mathcal{A}_t^n$ at time $t \in \mathcal{T}$, agent n's expected instantaneous utility at $h_t^n \in \mathcal{H}_t^n$ under μ is given by

$$\mathbb{E}_\mu^{g^{-n}} \left[\phi_t^n(C_t, X_t, A_t) | h_t^n, a_t^n\right]. \tag{D.22}$$

Since $A_t^k, k \neq n$ satisfies (5.27), the distribution of A_t^k only depends on X_t^k and B_t. Therefore, the distribution of A_t^{-n} only depends on X_t^{-n} and B_t. Then, for any realization $x_t^{-n} \in \mathcal{X}_t^{-n}$, $h_t^n = (x_{1:t}^n, h_t^c) \in \mathcal{H}_t^n$ and $a_t^n \in \mathcal{A}_t^n$,

$$\begin{aligned}
&\mathbb{E}_\mu^{g^{-n}} \left[\phi_t^n(C_t, X_t, A_t) | x_t^{-n}, a_t^n, h_t^n\right] \\
=&\mathbb{E}_\mu^{g^{-n}} \left[\phi_t^n(c_t, x_t, (a_t^n, A_t^{-n})) | x_t^{-n}, x_{1:t}^n, a_t^n, h_t^c, b_t\right] \\
=&\mathbb{E}^{g_t^{-n}} \left[\phi_t^n(c_t, x_t, (a_t^n, A_t^{-n})) | x_t^{-n}, b_t\right] \\
=:&\bar{\phi}_t^n(x_t, a_t^n, b_t, g_t^{-n});
\end{aligned} \tag{D.23}$$

the first equality in (D.23) holds because given ψ, $B_t = (C_t, \gamma_{\psi,t}(H_t^c), \hat{\gamma}_t(H_t^c))$ is a function of H_t^c; the second equality in (D.23) is true because the distribution of A_t^{-n} depends only on X_t^{-n}, B_t and the strategy g_t^{-n}. From (D.23), agent n's instantaneous utility (D.22) can be written as

$$\begin{aligned}
&\mathbb{E}_\mu^{g^{-n}} \left[\phi_t^n(C_t, X_t, A_t) | h_t^n, a_t^n\right] \\
=&\mathbb{E}_\mu^{g^{-n}} \left[\mathbb{E}_\mu^{g^{-n}} \left[\phi_t^n(C_t, X_t, A_t) | X_t^{-n}, a_t^n, h_t^n\right] | h_t^n, a_t^n\right] \\
=&\mathbb{E}_\mu^{g^{-n}} \left[\bar{\phi}_t^n((x_t^n, X_t^{-n}), a_t^n, b_t, g_t^{-n}) | h_t^c, x_{1:t}^n, a_t^n\right] \\
\overset{(a)}{=}&\mathbb{E}_\mu \left[\bar{\phi}_t^n((X_t^{-n}, x_t^n), a_t^n, b_t, g_t^{-n}) | h_t^c\right] \\
\overset{(b)}{=}&\mathbb{E}_{\pi_t} \left[\bar{\phi}_t^n((X_t^{-n}, x_t^n), a_t^n, b_t, g_t^{-n})\right] \\
=:&\tilde{\phi}_t^n(x_t^n, b_t, a_t^n, g_t^{-n}).
\end{aligned} \tag{D.24}$$

Equation (a) is true because, from Lemma V.14, X_t^{-n} and X_t^n are independent conditional on h_t^c. Equation (b) follows from the fact that π_t is the distribution of X_t conditional on h_t^c under μ, which is established by (5.32) in Lemma V.13.

Next, we show that the process $\{(X_t^n, B_t), t \in \mathcal{T}\}$ is a controlled Markov chain with respect to the action process $\{A_t^n, t \in \mathcal{T}\}$ for agent n.

From the system dynamics and the belief evolution (5.22), we know that (X_{t+1}^n, B_{t+1}) is a function of $\{X_t^n, Y_t^{-n}, A_t^n, A_t^{-n}, B_t, W_t\}$ where W_t denotes all the noises at time t. Furthermore, the distribution of (Y_t^k, A_t^k) depends only on $\{X_t^k, B_t, W_t, g_t^k\}$ for any $k \neq n$. Therefore,

$$(X_{t+1}^n, B_{t+1}) = \tilde{f}_t(X_t^n, X_t^{-n}, A_t^n, B_t, W_t, g_t^{-n}). \tag{D.25}$$

Suppose agent n uses an arbitrary strategy $\tilde{g}^n$. Then, for any realizations $x_{t+1}^n \in \mathcal{X}_{t+1}^n$, $b_{t+1} = (c_{t+1}, \pi_{t+1}, \hat{\pi}_{t+1}) \in \mathcal{B}_{t+1}$, $h_t^n = (x_{1:t}^n, h_t^c) \in \mathcal{H}_t^n$ and $a_t^n \in \mathcal{A}_t^n$, we obtain

$$\begin{aligned}
&\mathbb{P}_\mu^{\tilde{g}^n, g^{-n}}(x_{t+1}^n, b_{t+1} | h_t^n, a_t^n) \\
=&\sum_{x_t^{-n} \in \mathcal{X}_t^{-n}} \mathbb{P}_\mu^{\tilde{g}^n, g^{-n}}(x_{t+1}^n, b_{t+1} | x_t^{-n}, h_t^n, a_t^n) \mathbb{P}_\mu^{\tilde{g}^n, g^{-n}}(x_t^{-n} | h_t^n, a_t^n) \\
=&\sum_{x_t^{-n} \in \mathcal{X}_t^{-n}} \mathbb{P}_\mu^{\tilde{g}^n, g^{-n}}(x_{t+1}^n, b_{t+1} | x_t^{-n}, h_t^n, a_t^n) \pi_t(x_t^{-n}) \\
=&\sum_{x_t^{-n} \in \mathcal{X}_t^{-n}} \mathbb{P}^{g_t^{-n}}(x_{t+1}^n, b_{t+1} | x_t^n, x_t^{-n}, b_t, a_t^n) \pi_t(x_t^{-n}) \\
=&\mathbb{P}_\mu^{g^{-n}}(x_{t+1}^n, b_{t+1} | x_t^n, b_t, a_t^n).
\end{aligned} \tag{D.26}$$

The second equality in (D.26) follows from Lemma V.14 and (5.32) in Lemma V.13. The third equality in (D.26) follows from (D.25). The last equality follows from the same arguments as the first through third equalities.

Equation (D.26) shows that the process $\{(X_t^n, B_t), t \in \mathcal{T}\}$ is a controlled Markov Chain with respect to the action process $\{A_t^n, t \in \mathcal{T}\}$ for agent n. This process along

with the instantaneous utility (D.24) define a MDP. From the theory of MDP (see [1, Chap. 6]), there is an optimal strategy of agent n that is of the form

$$\lambda'^n_t(x^n_t, b_t) = \lambda'^n_t(x^n_t, (c_t, \gamma_{\psi,t}(h^c_t), \hat{\gamma}_t(h^c_t))) \tag{D.27}$$

for all $h^n_t = (x^n_{1:t}, h^c_t) \in \mathcal{H}^n_t$ for all $t \in \mathcal{T}$. This completes the proof of Lemma V.16.

□

Proof of Theorem V.20. Suppose (λ^*, ψ^*) solves the dynamic program defined by (5.40)-(5.43). Let $V^n_t, n \in \mathcal{N}, t \in \mathcal{T}$, denote the value functions computed by (5.40) and (5.43) from (λ^*, ψ^*). Then ψ^* is consistent with λ^* from (5.42).

Let $(g^*, \mu^*) = f(\lambda^*, \psi^*)$ defined by Lemma V.13. Then μ^* is consistent with g^* because of Lemma V.13. Furthermore, for all $n \in \mathcal{N}, t \in \mathcal{T}$, $V^n_t(x^n_t, b_t)$ (where $b_t = (c_t, \gamma_{\psi^*,t}(h^c_t), \hat{\gamma}_t(h^c_t)))$ is agent n's expected continuation utility from time t on under μ^* at $h^n_t = (x^n_{1:t}, h^c_t)$ when agent n uses g^{*n} and all other agents use g^{*-n}.

If every agent $k \neq n$ uses the strategy g^{*k}, from Lemma V.16 we know that there is a best response g'^n, under the belief system μ^*, of agent n such that for all $t \in \mathcal{T}$

$$g'^n_t(h^n_t) = \lambda'^n_t(x^n_t, b_t) \tag{D.28}$$

for some CIB strategy λ'^n_t for all $h^n_t = (h^c_t, x^n_{1:t})$. Define a CIB strategy profile $\lambda' := (\lambda'^n, \lambda^{*-n})$.

Let $V'^n_t, n \in \mathcal{N}, t \in \mathcal{T}$, be the functions generated by (5.40) and (5.43) from (λ', ψ^*). Then $V'^n_t(x^n_t, b_t)$ (where $b_t = (c_t, \gamma_{\psi^*,t}(h^c_t), \hat{\gamma}_t(h^c_t)))$ is agent n's expected continuation utility from time t on under μ^* at $h^n_t = (x^n_{1:t}, h^c_t)$ when agent n uses g'^n and all other agents use g^{*-n}. Since g'^n is a best response, for all $n \in \mathcal{N}, t \in \mathcal{T}$ and

$h_t^n = (x_{1:t}^n, h_t^c) \in \mathcal{H}_t^n$ we must have

$$V_t'^n(x_t^n, b_t) \geq V_t^n(x_t^n, b_t). \tag{D.29}$$

On the other hand, $V_t'^n(x_t^n, b_t)$ is player n's expected utility in stage game $G_t(V_{t+1}, \psi_t^*, b_t)$ when player n uses $\lambda'^n|_{b_t}$, and other players use $\lambda^{*-n}|_{b_t}$. However, from (5.41), $V_t^n(x_t^n, b_t)$ is player n's maximum expected utility in stage game $G_t(V_{t+1}, \psi_t^*, b_t)$ when other players use $\lambda^{*-n}|_{b_t}$ because the strategy $\lambda_t^{*n}|_{b_t}$ is a best response for player n in the stage game. This means that for all $n \in \mathcal{N}, t \in \mathcal{T}$ and $b_t \in \mathcal{B}_t$

$$V_t^n(x_t^n, b_t) \geq V_t'^n(x_t^n, b_t). \tag{D.30}$$

Combining (D.29) and (D.29) we get

$$V_t^n(x_t^n, b_t) = V_t'^n(x_t^n, b_t). \tag{D.31}$$

Equation (D.31) implies that, at any time t, the strategy $g_{t:T}^{*n}$ gives agent n the maximum expected continuation utility from time t on under μ^*. This complete the proof that (g^*, μ^*) is a PBE. As a result, the pair (λ^*, ψ^*) forms a CIB-PBE of the dynamic game described in Section 5.2.

□

In order to prove Theorem V.22, we first prove the following lemma.

Lemma D.1. *In **Game M***

$$\hat{\pi}_{t+1}^n = \hat{\psi}_t^n(y_t^n, \hat{\pi}_t). \tag{D.32}$$

Proof of Lemma D.1. From Lemma V.8

$$\begin{aligned}\hat{\pi}^n_{t+1} =& \hat{\psi}^n_t(y^n_t, a_t, \hat{\pi}_t)(x^n_{t+1}) \\ =& \frac{\sum_{x^n_t \in \mathcal{X}^n_t} p^n_t(x^n_{t+1}; x^n_t, a_t) q^n_t(y^n_t; x^n_t, a_t) \hat{\pi}^n_t(x^n_t)}{\sum_{x'^n_t \in \mathcal{X}^n_t} q^n_t(y^n_t; x'^n_t, a_t) \hat{\pi}^n_t(x'^n_t)}. \end{aligned} \tag{D.33}$$

Since $p^n_t(x^n_{t+1}; x^n_t, a_t) = p^n_t(x^n_{t+1}; x^n_t)$ and $q^n_t(y^n_t; x^n_t, a_t) = q^n_t(y^n_t; x^n_t)$ in **Game M**, the assertion of the lemma holds. □

Lemma D.1 shows that in **Game M** the signaling-free beliefs do not depend on the actions.

We now prove Theorem V.22.

Proof of Theorem V.22. Consider a CIB update rule ψ^* given by

$$\psi^{n*}_t(y^n_t, a_t, b_t) = \hat{\psi}^n_t(y^n_t, \hat{\pi}_t). \tag{D.34}$$

Based on ψ^* defined by (D.34), we solve the dynamic program defined by (5.40)-(5.43) to get a CIB strategy profile λ^* and show that (λ^*, ψ^*) forms a CIB-PBE for **Game M**. Note that under the update rule ψ^* given by (D.34), we have for any n and t

$$\Pi^n_t = \hat{\Pi}^n_t \tag{D.35}$$

Therefore, in the following we will replace Π^n_t by $\hat{\Pi}^n_t$ and drop Π^n_t if both Π^n_t and $\hat{\Pi}^n_t$ are present.

The dynamic program for **Game M** can be solved by induction. We prove the following claim:

At any time t, there exists a CIB strategy λ^*_t that satisfies (5.41), and the value

functions $V_t^n, n \in \mathcal{N}$, generated by (5.40) and (5.43) from $(\lambda_{t:T}^*, \psi_{t:T}^*)$ satisfy

$$V_t^n(x_t^n, b_t) = \tilde{U}_t^n(c_t, \hat{\pi}_t) + \tilde{V}_t^n(x_t^n, c_{t_k+1}, \hat{\pi}_t) \tag{D.36}$$

for some functions $\tilde{U}_t^n(c_t, \hat{\pi}_t)$ and $\tilde{V}_t^n(x_t^n, c_{t_k+1}, \hat{\pi}_t)$ when $t_k + 1 \leq t \leq t_{k+1}$ for some $t_k \in \overline{\mathcal{T}}$.

The above claim holds at $t = T + 1$ since $V_{T+1}^n = 0, n \in \mathcal{N}$.

Suppose the claim is true at $t + 1$.

At time $t_k + 1 \leq t < t_{k+1}$ for some $t_k \in \overline{\mathcal{T}}$, $X_{t+1}^n = X_t^n$, Y_t = empty, and $C_{t+1} = (C_t, A_t) = (C_{t_k+1}, A_{t_k+1:t})$. Then because of (5.57), (D.34) and the induction hypothesis for (D.36), player n's utility in stage game $G_t(V_{t+1}, \psi_t^*, b_t)$ is equal to

$$\begin{aligned}
&U^n_{G_t(V_{t+1}, \psi_t^*, b_t)} \\
=&\phi_t^n(c_t, X_t^{-n}, A_t) + V_{t+1}^n(X_{t+1}^n, C_{t+1}, \hat{\psi}_t(\hat{\pi}_t, Y_t)) \\
=&\phi_t^n(c_t, X_t^{-n}, A_t) + \tilde{U}_{t+1}^n((c_t, A_t), \hat{\psi}_t(\hat{\pi}_t)) + \tilde{V}_{t+1}^n(X_t^n, c_{t_k+1}, \hat{\psi}_t(\hat{\pi}_t))
\end{aligned} \tag{D.37}$$

for any $b_t \in \mathcal{B}_t$ and $n \in \mathcal{N}$. Define

$$\tilde{\phi}_t^n(X_t^{-n}, A_t, b_t) := \phi_t^n(c_t, X_t^{-n}, A_t) + \tilde{U}_{t+1}^n((c_t, A_t), \hat{\psi}_t(\hat{\pi}_t)), \tag{D.38}$$

$$\tilde{V}_t^n(X_t^n, c_{t_k+1}, \hat{\pi}_t) := \tilde{V}_{t+1}^n(X_t^n, c_{t_k+1}, \hat{\psi}_t(\hat{\pi}_t)). \tag{D.39}$$

At $t = t_k$ for some $t_k \in \overline{\mathcal{T}}$, $X_{t_k+1}^n = f_{t_k}^n(X_{t_k}^n, W_{t_k}^{n,X})$, $Y_{t_k}^n = h_{t_k}^n(X_{t_k}^n, W_{t_k}^{n,Y})$ and $C_{t_k+1} = f_{t_k}^c(C_{t_{k-1}+1}, W_{t_k}^C)$. Then because of (5.57), (D.34) and the induction hypoth-

esis for (D.36), player n's utility in stage game $G_{t_k}(V_{t_k+1}, \psi^*_{t_k}, b_{t_k})$ is equal to

$$\begin{aligned}
&U^n_{G_{t_k}(V_{t_k+1},\psi^*_{t_k},b_{t_k})} \\
=&\phi^n_{t_k}(c_{t_k}, X^{-n}_{t_k}, A_{t_k}) + V^n_{t_k+1}(X^n_{t_k+1}, C_{t_k+1}, \hat{\psi}_t(\hat{\pi}_{t_k}, Y_{t_k})) \\
=&\phi^n_{t_k}(c_{t_k}, X^{-n}_{t_k}, A_{t_k}) + \tilde{U}^n_{t+1}(f^c_{t_k}(c_{t_{k-1}+1}, W^C_{t_k}), \hat{\psi}_t(\hat{\pi}_t, h_{t_k}(X_{t_k}, W^Y_{t_k}))) \\
&+\tilde{V}^n_{t+1}(f^n_{t_k}(X^n_{t_k}, W^{n,X}_{t_k}), f^c_{t_k}(c_{t_{k-1}+1}, W^C_{t_k}), \hat{\psi}_t(\hat{\pi}_t)) \qquad \text{(D.40)}
\end{aligned}$$

for any $b_t \in \mathcal{B}_{t_k}$ and $n \in \mathcal{N}$. Define

$$\tilde{\phi}^n_{t_k}(c_{t_k}, X^{-n}_{t_k}, A_{t_k}, \hat{\pi}_{t_k}) := \phi^n_{t_k}(c_{t_k}, X^{-n}_{t_k}, A_{t_k}), \qquad \text{(D.41)}$$

$$\begin{aligned}
&\tilde{V}^n_{t_k}(X^n_{t_k}, c_{t_{k-1}+1}, \hat{\pi}_{t_k}) := \mathbb{E}_{\hat{\pi}_{t_k}}\Big[\tilde{U}^n_{t+1}(f^c_{t_k}(c_{t_{k-1}+1}, W^C_{t_k}), \hat{\psi}_t(\hat{\pi}_t, h_{t_k}(X_{t_k}, W^Y_{t_k}))) \\
&+ \tilde{V}^n_{t+1}(f^n_{t_k}(X^n_{t_k}, W^{n,X}_{t_k}), f^c_{t_k}(c_{t_{k-1}+1}, W^C_{t_k}), \hat{\psi}_t(\hat{\pi}_{t_k}))|X_{t_k}\Big]. \qquad \text{(D.42)}
\end{aligned}$$

Therefore, for any t, because of (D.37)-(D.42) player n's expected utility conditional on (X_t, A_t) in stage game $G_t(V_{t+1}, \psi^*_t, b_t)$, for $b_t = (c_t, \hat{\pi}_t) \in \mathcal{B}_t$, is equal to

$$\mathbb{E}_{\hat{\pi}_t}\left[U^n_{G_t(V_{t+1},\psi^*_t,b_t)}|X_t, A_t\right] = \tilde{\phi}^n_t(c_t, X^{-n}_t, A_t, \hat{\pi}_t) + \tilde{V}^n_t(X^n_t, c_{t_k+1}, \hat{\pi}_t). \qquad \text{(D.43)}$$

when $t_k + 1 \leq t \leq t_{k+1}$ for $t_k \in \overline{\mathcal{T}}$.

Since the second term in (D.43) does not depend on the players' strategies, an equilibrium of the stage game $G_t(V_{t+1}, \psi^*_t, b_t)$ is also an equilibrium of the game $G'_t(V_{t+1}, \psi^*_t, b_t)$ where each player $n \in \mathcal{N}$ has utility

$$U^n_{G'_t(V_{t+1},\psi^*_t,b_t)} := \tilde{\phi}^n_t(c_t, X^{-n}_t, A_t, \hat{\pi}_t). \qquad \text{(D.44)}$$

For any $b_t \in \mathcal{B}_t$, since $\mathcal{A}^n_t$ is a finite set for any $n \in N$ in **Game M**, the game $G'_t(V_{t+1}, \psi^*_t, b_t)$ has at least one Bayesian Nash equilibrium $\tilde{\lambda}^*_t(b_t) = \{\tilde{\lambda}^{*n}_t(b_t) \in \Delta(\mathcal{A}^n_t), n \in \mathcal{N}\}$ (see [5, 6]). Define $\lambda^{*n}_t(x^n_t, b_t) := \tilde{\lambda}^{*n}_t(b_t)$ for all $x^n_t \in \mathcal{X}^n_t, n \in \mathcal{N}$.

Then, we get a CIB strategy $\lambda_t^* \in BNE_t(V_{t+1}, \psi_t^*)$ so that (5.41) is satisfied at t. Moreover, from (5.43),

$$\begin{aligned}
&V_t^n(x_t^n, b_t) \\
=&D_t^n(V_{t+1}, \lambda_t^*, \psi_t^*)(x_t^n, b_t) \\
=&\mathbb{E}_{\hat{\pi}_t}^{\lambda_t^*}\left[\tilde{\phi}_t^n(c_t, X_t^{-n}, A_t, \hat{\pi}_t) + \tilde{V}_t^n(X_t^n, c_{t_k+1}, \hat{\pi}_t)|x_t^n\right] \\
\overset{(a)}{=}&\mathbb{E}_{\hat{\pi}_t}^{\lambda_t^*}\left[\tilde{\phi}_t^n(c_t, X_t^{-n}, A_t, \hat{\pi}_t)\right] + \tilde{V}_t^n(x_t^n, c_{t_k+1}, \hat{\pi}_t) \\
=:&\tilde{U}_t^n(c_t, \hat{\pi}_t) + \tilde{V}_t^n(x_t^n, c_{t_k+1}, \hat{\pi}_t)
\end{aligned} \tag{D.45}$$

where (a) is true because A_t^n only depends on b_t using λ_t^{*n}, and X_t^{-n} and X_t^n are independent under $\hat{\pi}_t$. Then (D.36) is satisfied at t, and the the proof of the claim is complete.

As a result of the claim, we obtain a CIB strategy profile λ^* and a CIB update rule ψ^* such that (5.40), (5.41) and (5.43) are satisfied. It remains to show the consistency (5.42). Using the dynamics of **Game M** and the fact that $\lambda_t^{*n}(x_t^n, b_t) := \tilde{\lambda}_t^{*n}(b_t)$, we obtain

$$\begin{aligned}
&\frac{\sum_{x_t^n \in \mathcal{X}_t^n} p_t^n(x_{t+1}^n; x_t^n, a_t)\eta_t^n(x_t^n, y_t^n, a_t, b_t)\pi_t^n(x_t^n)}{\sum_{x_t'^n \in \mathcal{X}_t^n} \eta_t^n(x_t'^n, y_t^n, a_t, b_t)\pi_t^n(x_t'^n)} \\
=&\frac{\sum_{x_t^n \in \mathcal{X}_t^n} p_t^n(x_{t+1}^n; x_t^n) q_t^n(y_t^n; x_t^n)\tilde{\lambda}_t^{*n}(b_t)(a_t^n)\hat{\pi}_t^n(x_t^n)}{\sum_{x_t'^n \in \mathcal{X}_t^n} q_t^n(y_t^n; x_t'^n)\tilde{\lambda}_t^{*n}(b_t)(a_t^n)\hat{\pi}_t^n(x_t'^n)} \\
=&\frac{\sum_{x_t^n \in \mathcal{X}_t^n} p_t^n(x_{t+1}^n; x_t^n) q_t^n(y_t^n; x_t^n)\hat{\pi}_t^n(x_t^n)}{\sum_{x_t'^n \in \mathcal{X}_t^n} q_t^n(y_t^n; x_t'^n)\hat{\pi}_t^n(x_t'^n)} \\
=&\hat{\psi}_t^n(y_t^n, \hat{\pi}_t)(x_t^n) = \psi_t^{*n}(y_t^n, a_t, b_t)(x_t^n).
\end{aligned} \tag{D.46}$$

Thus, ψ_t^* satisfies (5.28), and ψ_t^* is consistent with λ_t^*. Therefore (5.42) holds.

Since (λ^*, ψ^*) solves the dynamic program defined by (5.40)-(5.43), it is a CIB-PBE according to Theorem V.20. □

BIBLIOGRAPHY

BIBLIOGRAPHY

[1] P. Kumar and P. Varaiya, *Stochastic Systems: Estimation Identification and Adaptive Control*. Prentice-Hall, Inc., 1986.

[2] S. Yüksel and T. Başar, *Stochastic Networked Control Systems: Stabilization and Optimization under Information Constraints*. Systems and Control: Foundations and Applications, Birkhäuser, 2013.

[3] Y.-C. Ho, "Team decision theory and information structures," *Proc. IEEE*, vol. 68, no. 6, pp. 644–654, 1980.

[4] D. M. Kreps and J. Sobel, "Signalling," *Handbook of game theory*, vol. 2, pp. 849–867, 1994.

[5] D. Fudenberg and J. Tirole, *Game theory. 1991*. Cambridge, Massachusetts, 1991.

[6] M. J. Osborne and A. Rubinstein, *A course in game theory*. MIT press, 1994.

[7] T. Basar and G. J. Olsder, *Dynamic noncooperative game theory*, vol. 200. SIAM, 1995.

[8] R. B. Myerson, *Game theory: Analysis of Conflict*. Harvard university press, 2013.

[9] J. Filar and K. Vrieze, *Competitive Markov decision processes*. Springer-Verlag New York, Inc., 1996.

[10] V. D. Blondel and J. N. Tsitsiklis, "A survey of computational complexity results in systems and control," *Automatica*, vol. 36, no. 9, pp. 1249–1274, 2000.

[11] Q. Zhao and B. M. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Process. Mag.*, vol. 24, no. 3, pp. 79–89, 2007.

[12] D. S. Bernstein, R. Givan, N. Immerman, and S. Zilberstein, "The complexity of decentralized control of markov decision processes," *Math. Oper. Res.*, vol. 27, no. 4, pp. 819–840, 2002.

[13] A. Nayyar, A. Mahajan, and D. Teneketzis, "Decentralized stochastic control with partial history sharing: A common information approach," *IEEE Trans. Autom. Control*, vol. 58, no. 7, pp. 1644–1658, 2013.

[14] A. Nayyar, A. Gupta, C. Langbort, and T. Basar, "Common information based Markov perfect equilibria for stochastic games with asymmetric information: Finite games," *IEEE Trans. Autom. Control*, vol. 59, pp. 555–570, March 2014.

[15] A. Mahajan and D. Teneketzis, "Optimal design of sequential real-time communication systems," *IEEE Trans. Inf. Theory*, vol. 55, no. 11, pp. 5317–5338, 2009.

[16] E. Maskin and J. Tirole, "Markov perfect equilibrium: I. observable actions," *J. Econ. Theory*, vol. 100, no. 2, pp. 191–219, 2001.

[17] S. Zamir, "Repeated games of incomplete information: Zero-sum," *Handbook of Game Theory*, vol. 1, pp. 109–154, 1992.

[18] F. Forges, "Repeated games of incomplete information: non-zero-sum," *Handbook of Game Theory*, vol. 1, pp. 109–154, 1992.

[19] R. J. Aumann, M. Maschler, and R. E. Stearns, *Repeated games with incomplete information*. MIT press, 1995.

[20] G. J. Mailath and L. Samuelson, *Repeated games and reputations*, vol. 2. Oxford university press Oxford, 2006.

[21] L. Li and J. Shamma, "L_P formulation of asymmetric zero-sum stochastic games," in *Proc. 53rd IEEE Conf. Decision and Control (CDC)*, pp. 7752–7757, 2014.

[22] J. Renault, "The value of Markov chain games with lack of information on one side," *Math. Oper. Res.*, vol. 31, no. 3, pp. 490–512, 2006.

[23] F. Gensbittel and J. Renault, "The value of Markov chain games with incomplete information on both sides," *Math. Oper. Res.*, 2015. (in print).

[24] J. Renault, "The value of repeated games with an informed controller," *Math. Oper. Res.*, vol. 37, no. 1, pp. 154–179, 2012.

[25] P. Cardaliaguet, C. Rainer, D. Rosenberg, and N. Vieille, "Markov games with frequent actions and incomplete information-the limit case," *Math. Oper. Res.*, 2015.

[26] A. Gupta, A. Nayyar, C. Langbort, and T. Basar, "Common information based Markov perfect equilibria for linear-Gaussian games with asymmetric information," *SIAM J. Control Optim.*, vol. 52, no. 5, pp. 3228–3260, 2014.

[27] S. Ahmad, M. Liu, T. Javidi, Q. Zhao, and B. Krishnamachari, "Optimality of myopic sensing in multichannel opportunistic access," *IEEE Trans. Inf. Theory*, vol. 55, no. 9, pp. 4040–4050, 2009.

[28] P. Whittle, "Restless bandits: Activity allocation in a changing world," *J. Appl. Probab.*, vol. 25, pp. 287–298, 1988.

[29] J. Gittins, K. Glazebrook, and R. Weber, *Multi-Armed Bandit Allocation Indices*. WileyBlackwell, 2011.

[30] Q. Zhao, L. Tong, A. Swami, and Y. Chen, "Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: A POMDP framework," *IEEE J. Sel. Areas Commun.*, vol. 25, no. 3, pp. 589–600, 2007.

[31] A. Marshall, I. Olkin, and B. Arnold, *Inequalities: theory of majorization and its applications*. Springer Verlag, 2010.

[32] Q. Zhao, B. Krishnamachari, and K. Liu, "On myopic sensing for multi-channel opportunistic access: Structure, optimality, and performance," *IEEE Trans. Wireless Commun.*, vol. 7, no. 12, pp. 5431–5440, 2008.

[33] T. Javidi, B. Krishnamachari, Q. Zhao, and M. Liu, "Optimality of myopic sensing in multi-channel opportunistic access," in *Proc. IEEE Int. Conf. Commun. (ICC)*, (Beijing, China), pp. 2107–2112, 2008.

[34] S. Ahmad and M. Liu, "Multi-channel opportunistic access: A case of restless bandits with multiple plays," in *Proc. 47th Allerton Conf. Commun. Control Comput. (Allerton)*, (Monticello, IL), pp. 1361–1368, 2009.

[35] J. Gittins, "Bandit processes and dynamic allocation indices," *J. R. Stat. Soc.*, vol. 41, no. 2, pp. 148–177, 1979.

[36] R. Weber and G. Weiss, "On an index policy for restless bandits," *J. Appl. Probab.*, vol. 27, pp. 637–648, 1990.

[37] J. Niño-Mora, "Dynamic priority allocation via restless bandit marginal productivity indices," *TOP*, vol. 15, no. 2, pp. 161–198, 2007.

[38] J. L. Ny, M. Dahleh, and E. Feron, "Multi-uav dynamic routing with partial observations using restless bandit allocation indices," in *Proc. American Control Conf. (ACC)*, (Seattle, WA), pp. 4220–4225, 2008.

[39] K. Liu and Q. Zhao, "Indexability of restless bandit problems and optimality of Whittle index for dynamic multichannel access," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5547–5567, 2010.

[40] C. Lott and D. Teneketzis, "On the optimality of an index rule in multi-channel allocation for single-hop mobile networks with multiple service classes," *Probab. Eng. Inf. Sci*, vol. 14, no. 3, pp. 259–297, 2000.

[41] N. Ehsan and M. Liu, "Server allocation with delayed state observation: Sufficient conditions for the optimality of an index policy," *IEEE Trans. Wireless Commun.*, vol. 8, no. 4, pp. 1693–1705, 2009.

[42] S. Guha, K. Munagala, and P. Shi, "Approximation algorithms for restless bandit problems," *J. ACM*, vol. 58, no. 1, p. 3, 2010.

[43] M. P. Van Oyen and D. Teneketzis, "Optimal stochastic scheduling of forest networks with switching penalties," *Adv. Appl. Probab.*, pp. 474–497, 1994.

[44] P. Varaiya, J. Walrand, and C. Buyukkoc, "Extensions of the multiarmed bandit problem: the discounted case," *IEEE Trans. Autom. Control*, vol. 30, no. 5, pp. 426–439, 1985.

[45] W. Winston, "Optimality of the shortest line discipline," *J. Appl. Probab.*, pp. 181–189, 1977.

[46] R. R. Weber, "On the optimal assignment of customers to parallel servers," *J. Appl. Probab.*, pp. 406–413, 1978.

[47] E. Davis, *Optimal control of arrivals to a two-server queueing system with separate queues*. PhD thesis, PhD dissertation, Program in Operations Research, North Carolina State University, Raleigh, NC, 1977.

[48] A. Ephremides, P. Varaiya, and J. Walrand, "A simple dynamic routing problem," *IEEE Trans. Autom. Control*, vol. 25, no. 4, pp. 690–693, 1980.

[49] W. Lin and P. Kumar, "Optimal control of a queueing system with two heterogeneous servers," *IEEE Trans. Autom. Control*, vol. 29, no. 8, pp. 696–703, 1984.

[50] B. Hajek, "Optimal control of two interacting service stations," *IEEE Trans. Autom. Control*, vol. 29, no. 6, pp. 491–499, 1984.

[51] W. Whitt, "Deciding which queue to join: Some counterexamples," *Oper. Res.*, vol. 34, no. 1, pp. 55–62, 1986.

[52] R. R. Weber and S. Stidham Jr, "Optimal control of service rates in networks of queues," *Adv. Appl. Probab.*, pp. 202–218, 1987.

[53] A. Hordijk and G. Koole, "On the optimality of the generalized shortest queue policy.," *Probab. Eng. Inform. Sc.*, vol. 4, no. 4, pp. 477–487, 1990.

[54] A. Hordijk and G. Koole, "On the assignment of customers to parallel queues," *Probab. Eng. Inform. Sc.*, vol. 6, no. 04, pp. 495–511, 1992.

[55] R. Menich and R. F. Serfozo, "Optimality of routing and servicing in dependent parallel processing systems," *Queueing Syst.*, vol. 9, no. 4, pp. 403–418, 1991.

[56] R. D. Foley and D. McDonald, "Join the shortest queue: stability and exact asymptotics," *Ann. Appl. Probab.*, pp. 569–607, 2001.

[57] O. T. Akgun, R. Righter, and R. Wolff, "Understanding the marginal impact of customer flexibility," *Queueing Syst.*, vol. 71, no. 1-2, pp. 5–23, 2012.

[58] F. J. Beutler and D. Teneketzis, "Routing in queueing networks under imperfect information: Stochastic dominance and thresholds," *Stoch. Stoch. Rep.*, vol. 26, no. 2, pp. 81–100, 1989.

[59] J. Kuri and A. Kumar, "Optimal control of arrivals to queues with delayed queue length information," *IEEE Trans. Autom. Control*, vol. 40, no. 8, pp. 1444–1450, 1995.

[60] R. Cogill, M. Rotkowitz, B. Van Roy, and S. Lall, "An approximate dynamic programming approach to decentralized control of stochastic systems," in *Control of Uncertain Systems: Modelling, Approximation, and Design*, pp. 243–256, Springer, 2006.

[61] L. Ying and S. Shakkottai, "On throughput optimality with delayed network-state information," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5116–5132, 2011.

[62] A. A. Reddy, S. Banerjee, A. Gopalan, S. Shakkottai, and L. Ying, "On distributed scheduling with heterogeneously delayed network-state information," *Queueing Syst.*, vol. 72, no. 3-4, pp. 193–218, 2012.

[63] S. Manfredi, "Decentralized queue balancing and differentiated service scheme based on cooperative control concept," *IEEE Trans. Ind. Informat.*, vol. 10, pp. 586–593, Feb 2014.

[64] F. Abdollahi and K. Khorasani, "A novel H_∞ control strategy for design of a robust dynamic routing algorithm in traffic networks," *IEEE J. Sel. Areas Commun.*, vol. 26, no. 4, pp. 706–718, 2008.

[65] X. Si, X.-L. Zhu, X. Du, and X. Xie, "A decentralized routing control scheme for data communication networks," *Math. Probl. Eng.*, vol. 2013, 2013. Article ID 648267.

[66] D. G. Pandelis and D. Teneketzis, "A simple load balancing problem with decentralized information," *Math. Method Oper. Res.*, vol. 44, no. 1, pp. 97–113, 1996.

[67] A. Mahajan, "Optimal decentralized control of coupled subsystems with control sharing," *IEEE Trans. Autom. Control*, vol. 58, no. 9, pp. 2377–2382, 2013.

[68] R. J. Aumann, "Agreeing to disagree," *Ann. Stat.*, pp. 1236–1239, 1976.

[69] P. Bremaud, *Markov chains: Gibbs fields, Monte Carlo simulation, and queues*, vol. 31. springer, 1999.

[70] H. S. Witsenhausen, "Separation of estimation and control for discrete time systems," *Proc. IEEE*, vol. 59, no. 11, pp. 1557–1566, 1971.

[71] A. Nayyar, A. Mahajan, and D. Teneketzis, "The common-information approach to decentralized stochastic control," in *Information and Control in Networks*, pp. 123–156, Springer, 2014.

[72] R. Rom, M. Sidi, and R. R. M. Sidi, *Multiple Access Protocols: Performance and Analysis*. Springer-Verlag, 1990.

[73] G. I. Papadimitriou and A. S. Pomportsis, "Adaptive mac protocols for broadcast networks with bursty traffic," *IEEE Trans. Commun.*, vol. 51, no. 4, pp. 553–557, 2003.

[74] G. I. Papadimitriou, "A high performance MAC protocol for broadcast LANs with bursty traffic," *Comput. Commun.*, vol. 29, no. 8, pp. 994–997, 2006.

[75] J. Håstad, T. Leighton, and B. Rogoff, "Analysis of backoff protocols for multiple access channels," *SIAM J. Comput.*, vol. 25, no. 4, pp. 740–774, 1996.

[76] D. Shah, J. Shin, and P. Tetali, "Medium access using queues," in *52nd Ann. IEEE Symp. Found. (FOCS)*, pp. 698–707, IEEE, 2011.

[77] S. Rajagopalan, D. Shah, and J. Shin, "Network adiabatic theorem: an efficient randomized protocol for contention resolution," *ACM SIGMETRICS Performance Evaluation Rev. (PER)*, vol. 37, no. 1, pp. 133–144, 2009.

[78] D. Shah and J. Shin, "Randomized scheduling algorithm for queueing networks," *Ann. Appl. Probab.*, vol. 22, no. 1, pp. 128–171, 2012.

[79] L. Jiang, D. Shah, J. Shin, and J. Walrand, "Distributed random access algorithm: scheduling and congestion control," *IEEE Trans. Inf. Theory*, vol. 56, no. 12, pp. 6182–6207, 2010.

[80] L. Jiang and J. Walrand, "Approaching throughput-optimality in distributed CSMA scheduling algorithms with collisions," *IEEE/ACM Trans. Netw.*, vol. 19, no. 3, pp. 816–829, 2011.

[81] J. Ni, B. Tan, and R. Srikant, "Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks," *IEEE/ACM Trans. Netw.*, vol. 20, no. 3, pp. 825–836, 2012.

[82] J. Ghaderi Dehkordi, *Fundamental limits of random access in wireless networks*. PhD thesis, University of Illinois at Urbana-Champaign, 2013.

[83] L. Jiang, M. Leconte, J. Ni, R. Srikant, and J. Walrand, "Fast mixing of parallel Glauber dynamics and low-delay CSMA scheduling," *IEEE Trans. Inf. Theory*, vol. 58, no. 10, pp. 6541–6555, 2012.

[84] D. Lee, D. Yun, J. Shin, Y. Yi, and S.-Y. Yun, "Provable per-link delay-optimal CSMA for general wireless network topology," in *IEEE INFOCOM 2014*, pp. 2535–2543, 2014.

[85] S.-Y. Yun, Y. Yi, J. Shin, *et al.*, "Optimal CSMA: A survey," in *2012 IEEE Int. Conf. Commun. Syst. (ICCS)*, pp. 199–204, IEEE, 2012.

[86] B. S. Chlebus, D. R. Kowalski, and M. A. Rokicki, "Adversarial queuing on the multiple access channel," *ACM Trans. Algorithms*, vol. 8, no. 1, p. 5, 2012.

[87] L. Anantharamu, B. S. Chlebus, and M. A. Rokicki, "Adversarial multiple access channel with individual injection rates," in *Princ. Distrib. Syst.*, pp. 174–188, Springer, 2009.

[88] Q. Wang, *Optimal Channel-Switching Strategies in Multi-channel Wireless Networks*. PhD thesis, The University of Michigan, 2014.

[89] J. Luo and A. Ephremides, "On the throughput, capacity, and stability regions of random multiple access," *IEEE/ACM Trans. Netw.*, vol. 14, no. SI, pp. 2593–2607, 2006.

[90] W. Szpankowski, "Stability conditions for some distributed systems: buffered random access systems," *Adv. Appl. Probab.*, pp. 498–515, 1994.

[91] W. Luo and A. Ephremides, "Stability of N interacting queues in random-access systems," *IEEE Trans. Inf. Theory*, vol. 45, no. 5, pp. 1579–1587, 1999.

[92] S. Borst, M. Jonckheere, and L. Leskelä, "Stability of parallel queueing systems with coupled service rates," *Discrete Event Dyn. Syst.*, vol. 18, no. 4, pp. 447–472, 2008.

[93] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Cambridge University Press, 2009.

[94] D. Bertsekas and R. Gallager, *Data Networks*. Prentice-Hall international editions, Prentice Hall, 1992.

[95] S. P. Meyn and R. Tweedie, "Stability of Markovian processes I: criteria for discrete-time chains," *Adv. Appl. Probab.*, pp. 542–574, 1992.

[96] T. Borgers, D. Krahmer, and R. Strausz, *An introduction to the theory of mechanism design*. Oxford University Press, 2015.

[97] V. P. Crawford and J. Sobel, "Strategic information transmission," *Econometrica*, pp. 1431–1451, 1982.

[98] A. Nayyar and D. Teneketzis, "Sequential problems in decentralized detection with communication," *IEEE Trans. Inf. Theory*, vol. 57, no. 8, pp. 5410–5435, 2011.

[99] A. Nayyar and D. Teneketzis, "Signaling in sensor networks for sequential detection," *IEEE Trans. Control Netw. Syst.*, vol. 2, no. 1, pp. 36–46, 2015.

[100] Y. Ouyang and D. Teneketzis, "Signaling for decentralized routing in a queueing network," *Ann. Oper. Res.*, pp. 1–39, 2015.

[101] Y. Ouyang and D. Teneketzis, "A common information-based multiple access protocol achieving full throughput," in *Proc. 2015 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 31–35, 2015.

[102] H. Witsenhausen, "The intrinsic model for discrete stochastic control: Some open problems," in *Control Theory, Numerical Methods and Computer Systems Modelling*, pp. 322–335, Springer, 1975.

[103] H. S. Witsenhausen, "Equivalent stochastic control problems," *Math. Control Signal*, vol. 1, no. 1, pp. 3–11, 1988.

[104] Y. Ouyang, H. Tavafoghi, and D. Teneketzis, "Dynamic oligopoly games with private Markovian dynamics," in *Proc. 54th IEEE Conf. Decision and Control (CDC)*, 2015. to appear.

[105] K. E. Petersen and K. Petersen, *Ergodic theory*, vol. 2. Cambridge University Press, 1989.